Learning from Few Subjects with Large Amounts of Voice Monitoring Data

Jose Javier Gonzalez Ortiz[1] JJGO@MIT.EDU

Daryush D. Mehta[2] MEHTA.DARYUSH@MGH.HARVARD.EDU

Jarrad H. Van Stan[2] JVANSTAN@MGH.HARVARD.EDU

Robert Hillman[2] HILLMAN.ROBERT@MGH.HARVARD.EDU

John V. Guttag[1] GUTTAG@MIT.EDU

Marzyeh Ghassemi[3] MARZYEH@CS.TORONTO.EDU

1. *Computer Science and Artificial Intelligence Lab, MIT*
2. *Center for Laryngeal Surgery and Voice Rehabilitation, MGH*
3. *University of Toronto, Vector Institute*

Abstract

Recently, researchers have started training high complexity machine learning models to clinical tasks, often improving upon previous benchmarks. However, more often than not, these methods require large amounts of supervision to provide good generalization guarantees. When applied to data coming from small cohorts and long monitoring periods these models are prone to overfit to subject-identifying features. Since obtaining large amounts of labels is usually not practical in many scenarios, expert-driven knowledge of the task is a common technique to prevent overfitting. We present a two-step learning approach that is able to generalize under these circumstances when applied to a voice monitoring dataset. Our approach decouples the feature learning stage and performs it in an unsupervised manner, removing the need for laborious feature engineering. We show the effectiveness of our proposed model on two voice monitoring related tasks. We evaluate the extracted features for classifying between patients with vocal fold nodules and controls. We also demonstrate that the features capture pathology relevant information by showing that models trained on them are more accurate predicting vocal use for patients than for controls. Our proposed method is able to generalize to unseen subjects and across learning tasks while matching state-of-the-art results.

1. Introduction

Data regimes with a small number of subjects and large amounts of data per subject are common in many healthcare domains. Pathologies with low incidence rates can result in small patient cohorts. This is also the case when performing invasive monitoring or when specialized medical equipment is required. Similarly, long time series are commonplace in healthcare applications that require monitoring for extended periods as it is the case with sleep disorders.

Recent machine learning developments have led to significant improvements in classification accuracy for many clinical tasks (Rajpurkar et al., 2017; Henry et al., 2015; Poplin et al., 2018). Most of these approaches only work when vast amounts of data are available because they require large quantities of positive and negative training examples to provide good generalization guarantees. This often translates into needing either a large sample of patients and controls, or obtaining many labeled instances per subject.

In this work, we propose a two step framework that is able to generalize in the presence of few subjects but large amounts of data per subject. We first compute a general purpose time-frequency representation of the time series data. We then obtain feature encodings by training a deep convolutional autoencoder over this spectral information. We then use the encodings along with per-subject labels for downstream learning tasks. By decoupling the feature learning task from the limited supervision, we encourage the model to learn pathology related invariants rather than subject identifying characteristics.

We demonstrate the utility of our approach by applying it to a large collection of ambulatory voice monitoring data (Mehta et al., 2012). The dataset consists of 104 patients and controls, with each having multiple days of data ($\approx 10^9$ samples per subject). We compare to previous work (Ghassemi et al., 2014) which derived features using expert domain knowledge along with statistical aggregates to prevent overfitting. We show that training high complexity models on the soft per-subject labels leads to overfitting to subject-specific traits and fails to generalize to unseen subjects. In contrast, our proposed approach matches state-of-the-art predictive results without the need of laborious feature engineering.

We then evaluate the extracted features in a different task in the same dataset. We train a model to predict recent vocal load, i.e. the amount of recent voice usage, based on a short sample of consecutive encodings. We show the learned features capture pathology relevant information by analyzing the increase in model performance between patients with vocal fold nodules and their matched controls.

Technical Significance we present a learning based feature extraction model suitable for tasks for which there is a small number of subjects and large amounts of time series data per subject. We compare our model to baselines that are trained with direct supervision, and show a failure mode these models have. Under small patient cohorts and without fine grained supervision, fully supervised approaches can end up learning *subject-identifying* features instead of *pathology-related* features. To the best of our knowledge, the method we propose is the first for unsupervised feature extraction of large amounts of voice monitoring data with a small patient cohort.

Clinical Relevance Our proposed model aims to remove the need for laborious feature engineering. Even though the presented work is only evaluated in the context of voice monitoring data, we propose a methodology that can be applied to other tasks that fall in similar data regimes. Overall, this work represents a starting point on which others can build. In particular, we hypothesize that better techniques for dealing with ambulatory health related data could lead to further improvements in non-invasive and remote diagnostics.

In the specific context of voice monitoring, the proposed model is clinically relevant for several reasons. Vocal nodules are believed to be caused by damaging patterns of voice use, but the actual role of voice use in the etiology of vocal nodules is not well understood. The ability to detect the daily voice use patterns associated with vocal nodules (based on ambulatory monitoring) is an important step in developing improved methods for preventing, diagnosing, and treating this common disorder - including the potential use of this information in designing new ambulatory biofeedback approaches that could be used to more quickly modify and ameliorate damaging vocal behaviors.

2. Related Work

2.1. Ambulatory Medical Data

Ambulatory data collection techniques offer great potential for improving clinical care. For example, ambulatory cardiac monitoring techniques have been shown to be useful in the detection of hypertension (Verdecchia et al., 1994), atrial fibrillation (Jabaudon et al., 2004) and cardiac arrhythmias (Steinberg et al., 2017). Accelerometer data collected outside the clinical environment has been used for detecting physical activity and fall-detection systems (Mannini and Sabatini, 2010; Yuwono et al., 2012). This data regime is also frequent in the sleep analysis domain where patients need to be monitored for extended periods of time (Amiriparian et al., 2017; Biswal et al., 2017). Some recent work has leveraged ambulatory data collected from increasingly ubiquitous wearable devices to learn multiple medical conditions simultaneously (Ballinger et al., 2018).

The voice monitoring dataset we evaluate on has been previously used to distinguish between patients with vocal fold nodules and their associated controls (Ghassemi et al., 2014). This work relied on expert-driven features that prevented the models from overfitting to subjects.

2.2. Feature Extraction

Spectrograms are used extensively in the fields of music, navigational acoustics, and speech processing (Flanagan, 2013). Within the sound processing literature we find a variation: mel-frequency spectrograms (Imai, 1983). Values of the representation correspond to the logarithm of the power spectral density for different points in time and frequency. Values themselves are equally spaced in time and logarithmically scaled in frequency. Features computed in mel frequency are commonly used in speech recognition systems (Murty and Yegnanarayana, 2006; Ganchev et al., 2005). Mel frequency spectrograms have proven to be an effective representation for large-scale audio classification tasks using deep convolutional models (Hershey et al., 2017; Salamon and Bello, 2017). Recent work has shown the effectiveness of using mel spectrograms for training speech synthesis models (Shen et al., 2018).

Autoencoders have been previously proposed as a way to learn useful low feature representations of the data (Hinton and Salakhutdinov, 2006; Vincent et al., 2008). In the medical domain, unsupervised training of autoencoders has been successfully used in feature extraction task for time series data. They have been applied to electrocardiogram data (Al Rahhal et al., 2016), electroencephalogram data (Li et al., 2015) and polysomnogram data (Tsinalis et al., 2016). Similarly, researchers have been able to use autoencoder networks to learn from large amounts of wearable sensor data (Ballinger et al., 2018). These

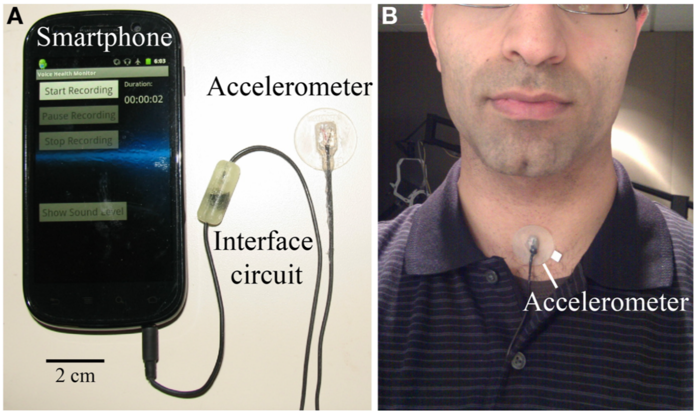


Figure 1: Ambulatory voice health monitor: (A) smartphone, accelerometer sensor, and cable with interface circuit encased in epoxy; (B) the wired accelerometer mounted on a pad affixed to the neck between the adam's apple and V-shaped notch of the collarbone.

approaches often rely on a large population size which is not a common case in many medical applications where specialized monitoring equipment is needed.

3. Data

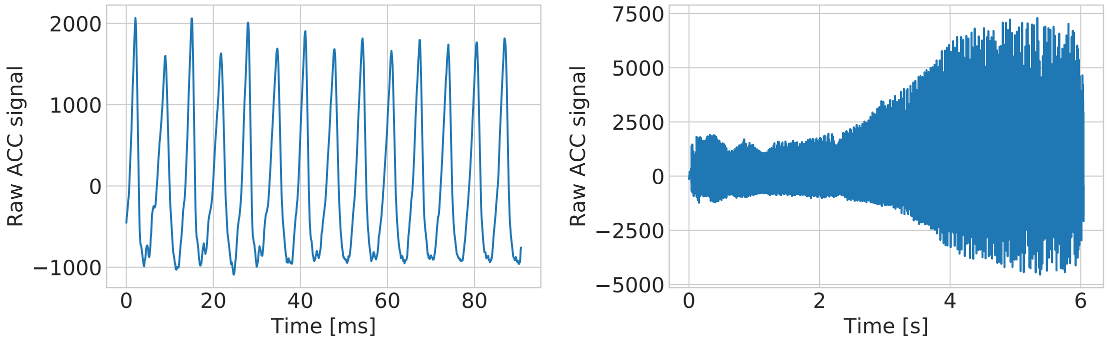
3.1. Data Extraction

Data was collected using an unobtrusive non-invasive ambulatory voice monitoring system that uses a neck-placed miniature accelerometer (ACC) as the phonation sensor and a smartphone as the data acquisition platform (Mehta et al., 2012). This device collects the unprocessed accelerometer signal and daily calibration recordings from speakers. The raw accelerometer signal is collected at an 11 025 Hz sampling rate, 16-bit quantization, and 80 dB dynamic range to get frequency content of neck surface vibrations up to 5000 Hz. Figure 1 depicts the ambulatory voice health monitor.

Accelerometer data is preferable to acoustic recordings for various reasons: 1) continuous daily recording of the acoustic signal raises privacy concerns, 2) the ACC signal is less affected by external acoustic noise sources (Zañartu et al., 2009), and 3) the ACC signal captured below the larynx is easier to analyze than the oral signal because the resonances of the respiratory system are relatively time-invariant compared to the vocal tract resonances.

All subjects were monitored over the course of at least one week using the described sensors. The subjects were instructed to wear the device during all waking hours. Nevertheless, data was not always acquired in an exhaustive or continuous fashion because of limitations of the data collection regime; strict compliance was not a pre-condition for data inclusion. For example, if a subject wore the device for only four hours on one day, we did not exclude data from that day from analysis.

Figure 2 shows samples of the raw accelerometer signal in both a short and a longer time scale. At a short time scale we can appreciate the individual glottal pulses, which have a fundamental frequency around 150 Hz. When looking at the longer time scale we see



(a)

(b)

Figure 2: Raw signal from the accelerometer. The zoomed in subfigure (a) shows the high frequency glottal segments. The zoomed out subfigure (b) showcases the signal envelope along with the ramp up related to the start in phonation.

the signal envelope responsible for the signal modulation along with the ramp up provoked from the start in phonation.

3.2. Cohort Selection

The collected dataset (Mehta et al., 2012) comprises 104 subjects that were monitored for roughly a week using the neck place accelerometer and a associated smartphone where the data is recorded. The population has 52 phonotraumatic patients with vocal fold lesions and 52 matched controls that are considered healthy speakers. Each patient typically aids in identifying a work colleague of the same gender and approximate age (±5 years) who has a normal voice. The normal vocal status of all control subjects is verified via interview and a laryngeal stroboscopic examination. Table 1 present some aggregate statistics for recorded times along with the percentage of voicing time.

Group	#	Days	Hours	Samples (millions)	% Voiced
PVH	52	7.33 ± 1.10	86.72 ± 20.40	3441.97 ± 809.52	9.27 ± 2.54
Controls	52	7.69 ± 1.11	94.76 ± 15.76	3761.15 ± 625.74	8.35 ± 2.98

Table 1: Mean and standard deviations across several metrics for both groups: Phonotraumatic Vocal Hyperfunction (PVH) and the matched controls. There are no statistically significant differences across any of these metrics.

3.3. Voicing Detection

For most tasks we want to ignore silent periods of time since 1) the pathology will not be manifested and 2) it will comprise the vast majority of the dataset. We do not have fine grained supervision of the voice monitoring signal. However, detecting voice activity is a rather straightforward task since voicing is directly correlated with the spectral intensity of the signal. We use these values as vocal use detection proxies as in (Ghassemi et al., 2014).

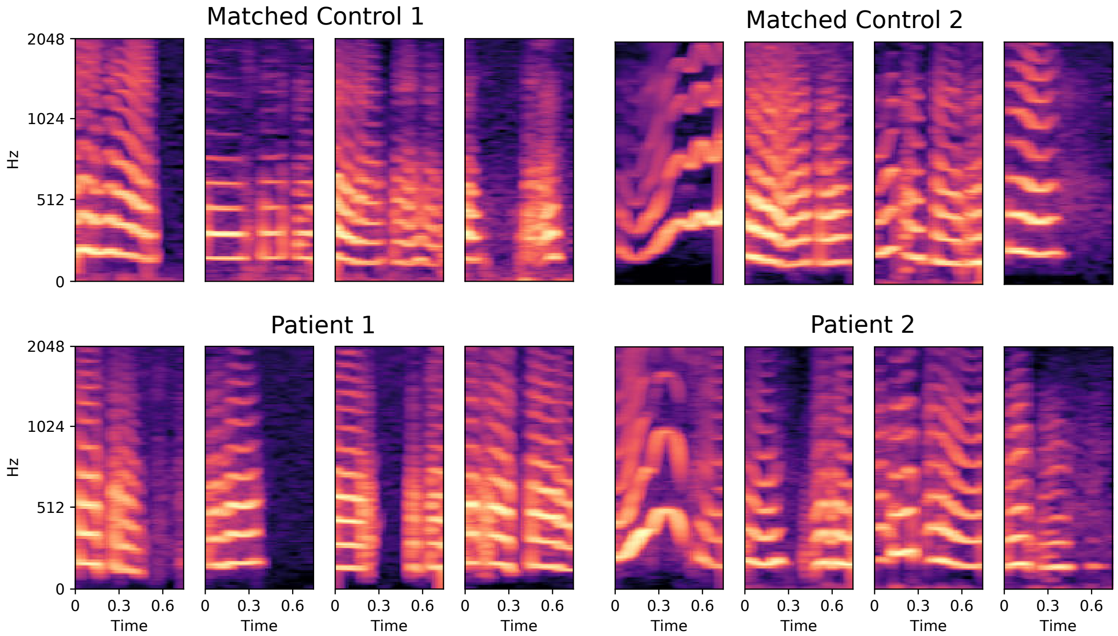


Figure 3: Four randomly sampled Log Mel-scaled Spectrograms for two patients with vocal hyperfunction and their corresponding matched controls. Both negative (top) and positive (bottom) windows can present similar patterns while still belonging to different classes, which makes the classification task challenging.

4. Methods

4.1. Time-Frequency Feature Representation

We aim to learn predictive features from the data distribution without learning subject-identifying patterns that could lead to overfitting. We are concerned with time series data since most ambulatory monitoring equipment collects data in this manner. We use tools from frequency domain analysis, common to science and engineering disciplines that have to work with time varying signals. Time-Frequency analysis transformations such as spectrograms convert a univariate signal from the time domain to a two dimensional time-frequency representation that contains the frequency domain transformation in a series of sliding windows.

For this work we do not use a raw spectrogram transformation. We use a mel-scaled spectrogram with logarithmic intensity as a two dimensional time-frequency encoding of the signal. Log mel frequency spectrograms have proven to be an effective representation for large-scale audio classification tasks using deep convolutional models (Hershey et al., 2017; Salamon and Bello, 2017). Values of the representation correspond to the logarithm of the power spectral density for different points in time and frequency, and values themselves are equally spaced in time and logarithmically scaled in frequency. We include some examples of this representation in Figure 3.

The bright bands in the spectrum correspond to the harmonics of the fundamental frequency of the speaker. We can verify that the fundamental frequency lies in the interval

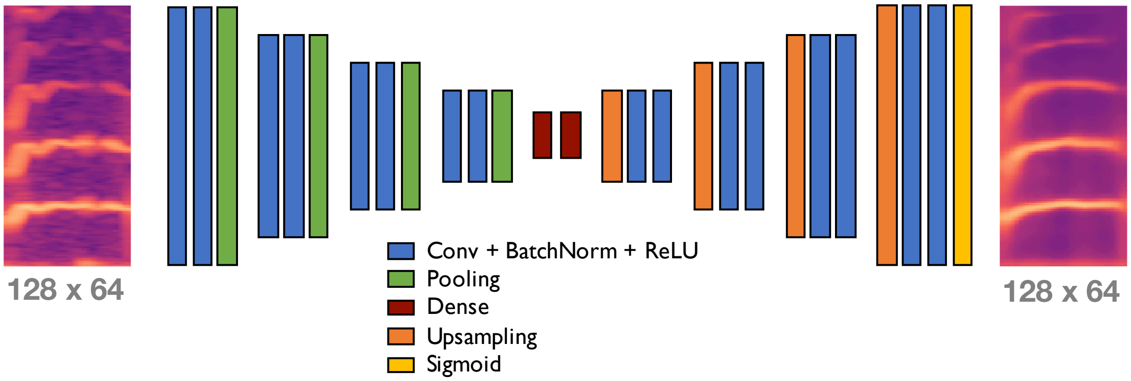


Figure 4: Diagram of the convolutional autoencoder architecture used to learn the low dimensional feature representations of the spectrogram data. Input and output are shown for a sample spectrogram. The network is able to retain the majority of the structure of the spectrogram representation.

150-200 Hz, common values for the human voice. From the figure we can appreciate the importance of using a logarithmic binning of frequencies. The uniform spacing in the bands is induced by this fact and ensures the representation has a higher resolution for low frequency phenomena. Similarly, computing the spectrum in a decibel scale is a common practice because of the multiplicative transformation most propagation channels induce.

4.2. Convolutional Autoencoder

From the mel-spectrogram representation we want a way to compress the information into a lower dimensionality embedding. There are many possible options for this kind of unsupervised feature extraction such as from PCA or clustering among others. Since the spectrogram is a two dimensional representation with significantly non-linear features (as shown in Figure 3) we chose a convolutional autoencoder trained in a self-supervised fashion. Convolutional neural networks have been shown to be able to encode highly non linear data distributions.

We train the model to output the same values provided as the input with a pixelwise mean squared error penalty, a common choice for a regression task as the one we have. Although we train the model to learn the identity function, a trivially looking task, it has the constraint to encode the representation in a low dimensional real valued vector as an intermediate step. This added constraint significantly increases the difficulty of the task and enforces the network to learn a compressed version of the input data, prioritizing encodings that will produce a better reconstruction of the output. We train the model on randomly sampled voicing segments of 0.74 s (the median voice segment length) with the start of the phonation aligned to be at the start of the segment.

5. Vocal Hyperfunction Classification

5.1. Experimental Setup

Prior work (Ghassemi et al., 2014) made use of statistical aggregates of expert-driven features to perform classification between patients with vocal fold nodules and their healthy matched controls. They learned a logistic regression model that predicted whether a voicing segment belonged to a patient or a control. Note that this is a form of soft label since it labels all voicing windows with the corresponding subject class. While this might be a reasonable thing to do in scenarios where the pathology is manifested uniformly throughout the data, it is often a simplifying assumption needed in cases where supervision is scarce. In our case, the belief is that control examples rarely manifest abnormal behavior. Some patient examples will manifest pathology relevant characteristics whereas others will correspond to normal instances of voicing activity.

Our proposed model like the one from prior work (Ghassemi et al., 2014), learns a mapping from the voicing segments to a binary value determining whether that segment belongs to a patient or a control. All the predicted labels from windows belonging to the same subject are then aggregated to produce a subject-level prediction. To aggregate predictions, we compute the percentage of windows labeled as positive for every subject and then choose a optimal threshold separating the two classes. Evaluation is then performed using the ground truth labels we have for every subject.

For the evaluation setup, we split the dataset into 5 randomized training/test splits. We use the first split of the data to perform the model selection and report the results for the remaining four. This is similar to a leave-one-out cross-validation strategy but less computationally expensive. The splits are stratified to maintain equal proportion on patients and controls, and to ensure that pairs of patient and matched control fall into the same split.

5.2. Benchmarks

For each set of experiments, we compare our proposed method to several benchmarks.

Feature-LR - As a first baseline method, we use an approach similar to (Ghassemi et al., 2014), which relies on expert-driven signal representations. The ACC signal is preprocessed by computing an array of features over 50 ms windows. For each window we compute three vocal dose measures: phonation time, cycle dose and distance dose. We also compute two general purpose signal processing features: sound pressure level and fundamental frequency. The features are then summarized using common statistical functions: mean, variance, skew, kurtosis and 5/95% percentiles. Then, we use the statistical aggregates to train an L1-regularized logistic regression model.

Feature-NN - We explore the use of sequence classification models to replace the aggregate measures. We train a 1-dimensional convolutional neural network and a GRU recurrent neural network (Cho et al., 2014) on sequences of the expert-driven features as input. We treat the specific sequence classifier implementation as a hyperparameter choice.We employ the same soft supervision as the Feature-LR approach, where each window was labeled with the subject class.

	Train		**Test**	
	AUROC	Accuracy	AUROC	Accuracy
Features-LR	0.70 (0.05)	0.71 (0.04)	0.68 (0.05)	0.69 (0.04)
Features-NN	0.90 (0.02)	0.91 (0.01)	0.50 (0.11)	0.48 (0.12)
Raw-NN	0.88 (0.04)	0.89 (0.04)	0.55 (0.11)	0.53 (0.07)
Ours	0.73 (0.06)	0.72 (0.04)	0.69 (0.07)	0.70 (0.05)

Table 2: Results for training and test results for the four splits of the data not used for model selection. Training values are included in order to quantify how much some models are overfitting. Mean (standard deviation) across the splits are reported. AUROC uses the continuous percentage output whereas accuracy employs the thresholded values.

Raw-NN - As an additional benchmark we train the same sequence classification models (CNN and GRU) with the raw accelerometer waveform using the same supervision approach. The main motivation for including this benchmark is because the preprocessed features the other benchmarks use could have subject-identifying properties.

5.3. Results

Table 2 reports the mean and standard deviation across the four splits not used for the hyperparameter selection. We observe that Feature-LR and our approach perform similarly in the training data, and each drops slightly in performance when presented with unseen data.

In contrast to those models, the Feature-NN and Raw-NN benchmarks strongly overfit, with extremely poor generalization results. This was true no matter what the choice of the neural network hyperparameters was. Furthermore, we experimented with various window sizes and feature subsets. Regardless of these choices, the model only improved in the training set while performing close to randomly on the unseen subjects. This demonstrates a dangerous failure mode of using large amounts of data and a small number of subjects with soft per-subject labels: fully supervised approaches can end up learning *subject-identifying* features instead of *pathology-related* features.

For the reported results, the learned encoding vectors had 20 dimensions, close to the 18 features of the statistical aggregates of Feature-LR. We chose this value to have comparable model complexities for these two classifiers. We did explore larger and smaller values of the size of the encoding vector. Smaller encoding vectors (5, 10, 15) performed worse than our results suggesting the model was underfitting. Larger values (30, 50, 70) did not increase or decrease performance significantly. As the size of the encoding got close to the size of the population (100, 150) the model started overfitting.

6. Voice Utilization Regression

6.1. Experimental Setup

To investigate the generality of our representation of the voice signal, we use the same representation for a different task, predicting recent voice utilization. Vocal loading refers to the stress the vocal folds experience when a person speaks. The current clinical understanding is that extended voicing causes more vocal loading for individuals with vocal fold nodules than for people with healthy vocal folds, and that this leads to short term changes in the activity of the vocal folds. We would like to test this hypothesis using ACC data.

Unfortunately, there is no accepted metric for quantifying vocal loading. What we do have is an unambiguous way of measuring voice utilization using the spectral intensity of the signal. Thus, we consider the problem of trying to predict the amount of recent vocalization from a sample of consecutive voiced segments. If it is easier to predict the amount of vocalization for patients than it is for controls, this would indicate that the amount of voicing has a greater effect on patients than on controls. It is plausible that this is, in turn, indicative of increased vocal loading, but there could be other causes.

In this experiment we want a mapping from several consecutive voicing segments to recent voicing activity. We compute voice usage labels using the same voicing detector we employed for segmenting the voiced windows. We generate labels by looking at each voicing window and computing the percentage of voicing time in the last ten minutes. We hypothesize that the model will produce more accurate results for patients than for the controls, since we expect their voice quality to be more affected.

To perform this experiment we train a regression model that takes a fixed size N of consecutive voicing windows and outputs a prediction of the amount of voicing in the previous ten minutes. Here N is treated as a hyperparameter of the model. Output labels are normalized to the length of the interval. We use a neural network model with a single hidden layer. The network is trained using a mean absolute error (MAE) cost function. We favor absolute error instead of squared error because of the presence of outliers in the label distribution.

6.2. Experimental Results

We train a model using 1,000 randomly sampled windows per patient. We split the population using the same strategy as in the classification task, using a single fold to perform the model selection. Out of considered values for N, $[1, 3, 10, 30]$, we found that a value of $N = 10$ performed best for the voice use regression task. We then compute the coefficient of the determination R^2 for each subject over the 1,000 predicted windows. In order to assess whether the model predictions are better for the patients than their matched controls we perform a paired t-test over the R^2 values for each patient and control (Fisher, 2006). As we previously described, for the considered population, each patient has a closely matched control with the same gender, and similar age and occupation.

We obtain that the difference is statistically significant with a p-value of $p = .04$. As we hypothesized, the model is more accurate when predicting recent voice usage for patients than for controls. We also highlight that this experiment was carried out using the same learned features without having to tweak them in any way for this particular task.

7. Conclusion

In this work, we proposed two step framework capable of learning useful features in a data regime with few subjects, little supervision and large amounts of time series per subject. Under this regime, traditional machine learning techniques are prone to overfitting issues by learning subject identifying features. We combine ideas from signal processing and unsupervised learning to learn a feature extraction model that works under these circumstances. We train a convolutional autoencoder on log mel-scaled spectrogram windows to extract features from the data. By decoupling the feature extraction from the downstream learning tasks, our learned representation prevents common overfitting issues that approaches with direct supervision experience.

We demonstrate the validity of our approach by applying it to a ambulatory voice monitoring dataset. First, we use the extracted features along with per-subject soft labels to classify between subjects with and without vocal fold nodules. Our framework generalizes well to unseen subjects and our results match the state-of-the-art performance on the classification task. We then use the learned features to predict recent voice usage based on a short sample of consecutive feature encodings. We show that the model is more accurate when predicting subjects with vocal fold nodules than when predicting their matched controls. Thus, the features generalize across subjects, while capturing relevant patterns for downstream clinical prediction tasks.

There are several directions for future work. First, in its current form our model requires small fixed sized windows of the time series data. While this is appropriate for datasets with many events during the course of a day, such as the voiced segments in ours, it does not translate well into problems with slowly varying behavior. A direction for future work is to explore how to extend or modify the proposed model to deal with this kind of data. As of now, the main limitation is that the model complexity increases linearly with the length of the signal. Future work could also explore how this approach can be applied to other datasets that operate in a similar data regime like for example ambulatory ECG data.

Acknowledgments

Jose Javier Gonzalez Ortiz is funded by “la Caixa” Foundation Fellowship. Dr. Marzyeh Ghassemi is funded in part by Microsoft Research, a CIFAR AI Chair at the Vector Institute, a Canada Research Council Chair, and an NSERC Discovery Grant. This research program is supported by the Voice Health Institute, the National Institutes of Health (NIH) National Institute on Deafness and Other Communication Disorders under Grant P50 02032.

References

Mohamad M Al Rahhal, Yakoub Bazi, Haikel AlHichri, Naif Alajlan, Farid Melgani, and Ronald R Yager. Deep learning approach for active classification of electrocardiogram signals. *Information Sciences*, 345:340–354, 2016.

Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Nicholas Cummins, Michael Freitag, Sergey Pugachevskiy, Alice Baird, and Björn W Schuller. Snore sound classification using image-based deep spectrum features. In *INTERSPEECH*, pages 3512–3516, 2017.

Brandon Ballinger, Johnson Hsieh, Avesh Singh, Nimit Sohoni, Jack Wang, Geoffrey H Tison, Gregory M Marcus, Jose M Sanchez, Carol Maguire, Jeffrey E Olgin, et al. Deepheart: Semi-supervised sequence learning for cardiovascular risk prediction. *arXiv preprint arXiv:1802.02511*, 2018.

Siddharth Biswal, Joshua Kulas, Haoqi Sun, Balaji Goparaju, M Brandon Westover, Matt T Bianchi, and Jimeng Sun. Sleepnet: automated sleep staging system via deep learning. *arXiv preprint arXiv:1707.08262*, 2017.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Ronald Aylmer Fisher. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 2006.

James L Flanagan. *Speech analysis synthesis and perception*, volume 3. Springer Science & Business Media, 2013.

Todor Ganchev, Nikos Fakotakis, and George Kokkinakis. Comparative evaluation of various mfcc implementations on the speaker verification task. In *Proceedings of the SPECOM*, number 2005 in 1, pages 191–194, 2005.

Marzyeh Ghassemi, Jarrad H Van Stan, Daryush D Mehta, Matías Zañartu, Harold A Cheyne II, Robert E Hillman, and John V Guttag. Learning to detect vocal hyperfunction from ambulatory neck-surface acceleration features: initial results for vocal fold nodules. *IEEE Trans. Biomed. Engineering*, 61(6):1668–1675, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (trewscore) for septic shock. *Science translational medicine*, 7 (299):299ra122–299ra122, 2015.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. CNN architectures for large-scale audio classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 131–135. IEEE, 2017.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Satoshi Imai. Cepstral analysis synthesis on the mel frequency scale. In *ICASSP'83. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 8, pages 93–96. IEEE, 1983.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 2015.

Denis Jabaudon, Juan Sztajzel, Katia Sievert, Theodor Landis, and Roman Sztajzel. Usefulness of ambulatory 7-day ecg monitoring for the detection of atrial fibrillation and flutter after acute stroke and transient ischemic attack. *Stroke*, 35(7):1647–1651, 2004.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Junhua Li, Zbigniew Struzik, Liqing Zhang, and Andrzej Cichocki. Feature learning from incomplete eeg with denoising autoencoder. *Neurocomputing*, 165:23–31, 2015.

Andrea Mannini and Angelo Maria Sabatini. Machine learning methods for classifying human physical activity from on-body accelerometers. *Sensors*, 10(2):1154–1175, 2010.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25, 2015.

Daryush D Mehta, Matias Zanartu, Shengran W Feng, Harold A Cheyne II, and Robert E Hillman. Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform. *IEEE Transactions on Biomedical Engineering*, 59(11):3090–3096, 2012.

K Sri Rama Murty and Bayya Yegnanarayana. Combining evidence from residual phase and mfcc features for speaker recognition. *IEEE signal processing letters*, 13(1):52–55, 2006.

Ryan Poplin, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158, 2018.

Pranav Rajpurkar, Awni Y Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y Ng. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836*, 2017.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3): 279–283, 2017.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Jonathan S Steinberg, Niraj Varma, Iwona Cygankiewicz, Peter Aziz, Paweł Balsam, Adrian Baranchuk, Daniel J Cantillon, Polychronis Dilaveris, Sergio J Dubner, Nabil El-Sherif, et al. 2017 ishne-hrs expert consensus statement on ambulatory ecg and external cardiac monitoring/telemetry. *Heart Rhythm*, 14(7):e55–e96, 2017.

Orestis Tsinalis, Paul M Matthews, and Yike Guo. Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. *Annals of biomedical engineering*, 44(5):1587–1597, 2016.

Paolo Verdecchia, Carlo Porcellati, Giuseppe Schillaci, Claudia Borgioni, Antonella Ciucci, Massimo Battistelli, Massimo Guerrieri, Camillo Gatteschi, Ivano Zampi, Antonella Santucci, et al. Ambulatory blood pressure. an independent predictor of prognosis in essential hypertension. *Hypertension*, 24(6):793–801, 1994.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1096–1103, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390294.

Mitchell Yuwono, Bruce D Moulton, Steven W Su, Branko G Celler, and Hung T Nguyen. Unsupervised machine-learning method for improving the performance of ambulatory fall-detection systems. *Biomedical engineering online*, 11(1):9, 2012.

Matías Zañartu, Julio C Ho, Steve S Kraman, Hans Pasterkamp, Jessica E Huber, and George R Wodicka. Air-borne and tissue-borne sensitivities of bioacoustic sensors used on the skin surface. *IEEE Transactions on Biomedical Engineering*, 56(2):443–451, 2009.

Appendix A. Implementation Details

We include details of the implementation to facilitate the reproduction of the experiments carried out in this work.

Spectrogram Computation As we have discussed, the main features of the spectrogram computation is a logarithmic binning of frequency along with having a log scale in the intensities to prevent spikes from dominating the entire signal.

We compute the power spectrum of the signal and ignore the phase content by taking its magnitude. We did this to ensure that the spectrogram was a real-valued and because of the modulated nature of the glottal pulse signal. We computed these spectrograms using $N_{\mathrm{NFFT}} = 2048$, $N_{\mathrm{filters}} = 128$, and a sliding window setup with distance of $\Delta_w = 128$ samples. The size of the voicing segments was chosen to be close to the length of the median voiced segment, 799 ms. Thus, we set the number of windows to $N_{\mathrm{w}} = 64$ since for a sampling frequency of $f_0 = 11\,025\,\mathrm{Hz}$ we get:

$$T = N_w \Delta_w \frac{1}{f_0} = 128 \cdot 64 \frac{1}{11025} \approx 743\,\mathrm{ms}$$

We favor 64 instead of a more accurate 68 since a power of two makes pooling and upsampling the data a simpler task in the convolutional neural network. This removes the need for padding and cropping, simplifying the pipeline for the neural network and making the computation more efficient. Once computed using the mentioned N_{NFFT} value, we band-limited the spectrograms from 8192 Hz to 2048 Hz since the majority of the energy content of the signal was present in that region.

Lastly, we mention that the choice of overlap in the sliding window along with the choice of window length was arbitrary (given the aforementioned constraints) and was not cross validated as part of the model. The reasoning behind this choice was to use a reasonable off-the-shelf time-frequency representation and then use the unsupervised model to perform the feature extraction. If we were to compute highly tuned spectrograms for the task at hand, the spectrogram itself would become an expert engineered set of features.

Autoencoder Model Following the same approach as with the time-frequency representation, we use an off-the-shelf convolutional autoencoder model with default choices for the majority of its settings. The aim is not to over-engineer the network to the voice monitoring dataset.

As pictured in Figure 4, the model is composed of a series of encoding blocks, an embedding block, a series of decoding blocks and a final output layer.

- **Encoding blocks** - these are composed of two consecutive two dimensional convolutional layers. We employed rectified linear unit (ReLU) activations because of their known empirical results on image classification and segmentation tasks (Simonyan and Zisserman, 2014; Ronneberger et al., 2015). Similarly, we use convolutional kernels of size 3 by 3 to limit the model complexity. We add a batch normalization step between the convolutional filters and the activation function (Ioffe and Szegedy, 2015). Batch Normalization proved to be useful in reducing the number of epochs until the network converged. Following the two convolutional layers, we add a max pooling layer that downsamples the images by a factor of two.

- **Embedding block** - we implement the embedding by using two dense layers, one that goes from the last encoding block to the embedding vector size and then a symmetric one that goes from said vector to the first decoding block. We experimented with various embedding vector sizes and ended up using 30 units for the reported results.
- **Decoding blocks** - these blocks contain two identical convolutional layers to those described in the encoding blocks. The maxpooling downsampling at the end is replaced by a upsampling that precedes the convolutional layers.
- **Output layer** - since the output of our task was a real valued signal with a single channel, we employed a sigmoid activation after the last decoding block.

We trained the described model with a mean squared error loss until convergence. All of the weights in the network were initialized with He-Normal (He et al., 2015) distributed values. The optimizer chosen for the training was the Adam (Kingma and Ba, 2014) optimizer with an initial learning rate of $\eta_0 = 10^{-3}$ and momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$. Both the input and output size were 128 by 64 pixel images, as discussed previously.

Software Libraries and Versions All the experiments were designed and executed using `Python 3.6.4` compiled against the `Anaconda` framework 4.4.10 for Intel Math Kernel Library Support. General tensor operations were carried out with `Numpy` 1.14.0 and the logistic regression models were trained using the `scikit-learn` library with version 0.19.1. For the melspectrograms computation we employed the `librosa` (McFee et al., 2015) module with version 0.6.1.

For the implementation of the deep neural network models, we used the `Keras` library with the `TensorFlow` backend configuration with respective versions 2.1.6 and 1.8.0. The GPU version of `TensorFlow` was used to speed up the experiment execution time. The `CUDA` driver library had version 9.0 and the `cuDNN` Deep Neural Network Library had version 7.0.