A MIXTURE OF EXPERTS APPROACH TOWARDS INTELLIGIBILITY CLASSIFICATION OF PATHOLOGICAL SPEECH

Rahul Gupta[1]*, Kartik Audhkhasi*[2]*, Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Lab (SAIL), Los Angeles, California, USA
[2] IBM T. J. Watson Research Center, Yorktown Heights, New York, USA

ABSTRACT

Pathological speech involves atypical speech production which may result from several factors including oral diseases, physical disabilities in the voice production system and atypical anatomy. Automatic evaluation of intelligibility in patients with pathological speech can assist accurate diagnosis of pathological conditions. Loss of intelligibility may be associated with one of the several pathological conditions, making automatic evaluation a challenging computational problem. A Mixture of Experts (MoE) models class boundaries using a weighted combination of several experts and can characterize the complex class boundaries arising due to pathological variability. We train an MoE for intelligibility evaluation using a modified Expectation Maximization (EM) algorithm based on joint simulated annealing-gradient ascent procedure. Our algorithm optimizes the expert parameters and simultaneously obtains the feature subsets for each expert. We observe that the MoE trained using the new EM algorithm not only outperforms a single classifier baseline but also the vanilla MoE. We perform further data analysis and interpret the weights assigned to each expert during inference. Also, we obtain a different feature subset per expert in the mixture. This illustrates feature use based on location of the data point in the feature space.

Index Terms— Pathological speech, intelligibility, Mixture of Experts, simulated annealing, gradient ascent

1. INTRODUCTION

Voice production in patients with pathological speech suffers from atypicality resulting due to factors including oral diseases [1], injury [2] and/or genetic anomaly [3]. One or more components of the voice production system are compromised (e.g. larynx, vocal chords) leading to loss of intelligibility. Automatic evaluation of intelligibility may aid diagnosis and provide a fair assessment of the pathological severity. However the heterogeneity in the causes of pathological speech [4] makes this a challenging computational problem. The sources of atypicality may impact the vocal aspects dissimilarly. For instance head tumors may lead to loss of motor control whereas a physical injury may cause atypical anatomy. The challenge lies not only in robustly modeling the variability in intelligibility characteristics but also providing sufficient interpretability for a more informed diagnosis. We use a Mixture of Experts (MoE) [5, 6] framework to address both these aspects. Inference in an MoE is performed based on a weighted combination of outputs from multiple experts. Use of multiple experts can robustly model the class boundaries over the feature space. Furthermore, the weight assigned to each expert may inform us about the class boundary and consequently the nature of pathological condition.

Studies have proposed several vocal features (prosodic and frequency spectrum based) for evaluation of pathological speech [7–9]. Equal emphasis has been laid in evaluating various machine learning tools such as Gaussian mixture models [10], neural networks [11] and wavelet packet decomposition [12, 13] for explaining characteristics of pathological speech. Middag et al. [14] designed regression models on phonological features to evaluate intelligibility in pathological speech. The Interspeech pathology sub-challenge 2012 [15] led to several investigations on intelligibility in pathological data. Under the same challenge, Kim et al. [16] and Huang et al. [17] applied fusion techniques on multiple systems to infer intelligibility. In another paper Huang et al. [18], propose asymmetric sparse kernel partial least squares classifier for the same problem. Despite providing a good understanding of the relation between intelligibility and vocal features, existing research uses machine learning tools which may lack robustness and/or interpretability. Simple classifiers fail to model complex class boundaries whereas strong classifiers may be less interpretable. We address these issues in this paper using an MoE framework. An MoE uses an ensemble of experts operating collectively over the feature space. Based on the data sample at hand, each expert is weighted differently to obtain the final outcome. An MoE can model the variability in class boundaries introduced due to heterogeneity in pathological conditions. Also, the expert weights for a data sample may inform about its proximity to other data samples and consequently about the pathological condition.

Furthermore, we propose a modified EM algorithm to jointly train and obtain the feature subset for each expert. The feature subsets per expert may inform us about the utility of a feature over the feature space. Feature selection involves binary integer programming (NP-hard) and is not trivial for an MoE. Wrapper (e.g. forward feature selection) and filter methods (e.g. correlation based selection) yield feature subsets but either take considerable time or select based on a heuristic based objective function. Previous studies [19–21] have proposed L1/L2 regularization to obtain the feature subset on an MoE. However, this leads to modification of the objective function. We address these issues by using a simulated annealing-gradient ascent based approach. We proposed a similar approach in [22] to sequentially train an ensemble of classifiers. However each expert was trained individually, unlike the joint optimization in an MoE.

In the next section we describe the database. Section 3 describes the experimental method. We list the results in section 4 followed by data analysis in section 5. We present the conclusions in section 6.

2. DATABASE

We use the NKI CCRT Speech Corpus [23] collected by the Department of Head and Neck Oncology and Surgery at the Netherlands Cancer Institute. We use recordings from a set of 55 speakers with available perceptual intelligibility ratings on a scale of 1-7. These ratings were obtained using majority voting on evaluations from thirteen speech pathologists. For our purposes, we binarize the ratings using same discretization criteria (median of scale) as was chosen

Feature	Statistical functionals
Pitch, intensity, jitter, RASTA-style auditory spectrum (bands 1-26), voicing probability	mean, range, quartiles, standard deviation, maximum, minimum

Table 1. List of features used in intelligibility classification from pathological speech.

in the Interspeech challenge 2012 [15] on the same dataset. We label the recordings as being intelligible (I) or non-intelligible (NI). Although this leads to a coarser estimation of high vs low intelligibility, we obtain more representative data per class thereby reducing data sparsity. Overall we have 2379 utterances (NI: 1181, I: 1198) with 51 utterances from most of the speakers. Note that the utterances from the same speaker are evaluated independently and may fall in either of the two classes. For the classification experiment, we extract several audio features as described next.

2.1. Acoustic-prosodic features

We use statistical functionals over several prosodic cues and RASTA-PLP based spectrum (extracted at 100 frames/second) per utterance as our features. Several studies have showed the effectiveness of similar spectral and prosodic features [9, 15]. The features are extracted using Opensmile [24] and the list of features is shown in Table 1. The RASTA-PLP based spectrum and prosodic signals were z-normalized per speaker. All inclusive, we obtain a feature set of dimensionality $D = 240$. We represent the feature vector from the n^{th} instance as $\boldsymbol{x}_n = [x_n^1, .., x_n^D]$. The class label y is drawn from the set {I, NI}.

3. EXPERIMENTAL METHOD

We use a Mixture of Experts (MoE) framework [5] to address the variability in pathological conditions. We expect each expert to capture the intelligibility class boundaries as dictated by the heterogeneous pathological conditions. Moreover, not all features may be locally useful in assessing intelligibility. We propose an extension to the vanilla MoE training scheme, performing joint feature selection for each classifier along with class boundary computation. We describe the MoE framework in detail below.

3.1. Mixture of Experts

The probability $p(y|\boldsymbol{x}_n)$ of y being the true class, given the feature $\boldsymbol{x}_n$ and a mixture of K experts is computed as shown in (1). m_k is a latent membership variable which determines if $\boldsymbol{x}_n$ is modeled by the expert k. $p(y, m_k|\boldsymbol{x}_n)$ represents the joint probability of y and m_k given the features. $p(y|\boldsymbol{x}_n)$ is obtained as a result of marginalizing $p(y, m_k|\boldsymbol{x}_n)$ over all the latent membership variables $\{m_1, .., m_K\}$. $p(y, m_k|\boldsymbol{x}_n)$ splits into two parts: (i) class probability from expert k and (ii) a gating function for expert k. We explain these two parts below.

$$p(y_n|\boldsymbol{x}_n) = \sum_{k=1}^{K} p(y_n, m_k|\boldsymbol{x}_n) = \sum_{k=1}^{K} \underbrace{p(y_n|m_k, \boldsymbol{x}_n)}_{\substack{\text{class probability} \\ \text{from expert } k}} \underbrace{p(m_k|\boldsymbol{x}_n)}_{\substack{\text{gating function} \\ \text{for expert } k}} \quad (1)$$

(i) *Class probability from expert k*: This part of the equation yields the probability of y given $\boldsymbol{x}_n$ as given by the expert k. We use logistic regression models as our experts. The class boundary for the expert k is determined based on the weight vectors $\boldsymbol{\theta}_{k,y}$ ($y \in$ {I, NI}). In a vanilla MoE, the weight vectors $\boldsymbol{\theta}_{k,y}$ operate on all the features. However in the proposed MoE architecture, these weight vectors operate on a subset of features determined by a D dimensional binary feature set vector $\boldsymbol{\lambda}_k = \{\lambda_k^1, .., \lambda_k^d, .., \lambda_k^D\}$. $\lambda_k^d = 1$ indicates the use of the d^{th} feature by the expert k. The probability $p(y|m_k, \boldsymbol{x}_n)$ is shown in (2). *Diag* $\boldsymbol{\lambda}_k$ is a diagonal matrix containing the binary variables λ_k^d. Note that a $\lambda_k^d = 0$ would render the d^{th} feature ineffective for expert k.

$$p(y|m_k, \boldsymbol{x}_n) = \frac{\exp\left((\boldsymbol{\theta}_{k,y})^T \times \mathit{Diag}(\boldsymbol{\lambda}_k) \times \boldsymbol{x}_n\right)}{\sum\limits_{y \in \{\text{I, NI}\}} \exp\left((\boldsymbol{\theta}_{k,y})^T \times \mathit{Diag}(\boldsymbol{\lambda}_k) \times \boldsymbol{x}_n\right)} \quad (2)$$

(ii) *Gating function for expert k*: The gating function $p(m_k|\boldsymbol{x}_n)$ weighs the class probability from expert k based on the features. A higher gating value implies a higher confidence in the expert. We use a sigmoid gating function as shown in (3). $\boldsymbol{\phi}_k$ is the gating vector for expert k. Note that we also can learn a feature subset for the gating function. However, we did not obtain a significant improvement in performance by incorporating it.

$$p(m_k|\boldsymbol{x}_n) = \frac{\exp\left((\boldsymbol{\phi}_k)^T \times \boldsymbol{x}_n\right)}{\sum_{k'=1}^{K} \exp\left((\boldsymbol{\phi}_{k'})^T \times \boldsymbol{x}_n\right)} \quad (3)$$

We expect each expert to reliably model the class boundaries in a part of the feature space. The final outcome is a weighted sum of the outputs from each expert based on the gating function. Despite receiving enough attention [25], determining a subset of features while training a classifier in not trivial. In particular, it is more challenging during joint optimization of classifier parameters in an MoE. We propose an extension to EM algorithm based MoE training to incorporate feature selection. Our algorithm performs a joint simulated annealing-gradient ascent to determine the weight vectors $\boldsymbol{\theta}_{k,y}$, gating vector $\boldsymbol{\phi}_k$ and the binary vector $\boldsymbol{\lambda}_k$. In the next section, we describe the modified EM algorithm.

3.1.1. Modified EM based on joint simulated annealing-gradient ascent

We optimize the data log-likelihood to determine the parameters $\boldsymbol{\theta}_{k,y}$, $\boldsymbol{\lambda}_k$ and $\boldsymbol{\phi}_k$. Given N data points $\langle \boldsymbol{x}_1, .., \boldsymbol{x}_n, .., \boldsymbol{x}_N \rangle$ and corresponding true labels $\langle y_1, .., y_n, .., y_N \rangle$ the data log-likelihood ($\mathcal{L}$) is:

$$\mathcal{L} = \sum_{n=1}^{N} \log p(y_n|\boldsymbol{x}_n) \quad (4)$$

For an MoE, $\mathcal{L}$ is equivalent to the function shown in (5) (for detailed derivation please refer to [26], section 3.2). R_{kn} is the responsibility of expert k for $\boldsymbol{x}_n$ as given by (6). A high R_{kn} implies that the expert k is weighted more while modeling $\boldsymbol{x}_n$.

$$\mathcal{L} = \sum_{n=1}^{N} \sum_{k=1}^{K} R_{kn} \big(\log p(y_n|\boldsymbol{x}_n, m_k) + \log p(m_k|\boldsymbol{x}_n) \big) \quad (5)$$

$$R_{kn} = \frac{p(y_n, m_k|\boldsymbol{x}_n)}{\sum_{k'=1}^{K} p(y_n, m_{k'}|\boldsymbol{x}_n)} \quad (6)$$

In an EM algorithm for a vanilla MoE, determining R_{kn} is the expectation step (E-step) and optimizing the parameters $\langle \boldsymbol{\theta}_{k,\text{I}}; \boldsymbol{\theta}_{k,\text{NI}}; \boldsymbol{\phi}_k \rangle$ with $\boldsymbol{\lambda}_k = \{1, ..(\text{D times}).., 1\}$ is the maximization step (M-step). In the modified EM algorithm, we change the M-step to consider an alternate feature set vector ($\boldsymbol{\lambda}_k^{(\text{ii})}$) in addition to the one already existing ($\boldsymbol{\lambda}_k = \boldsymbol{\lambda}_k^{(\text{i})}$). $\boldsymbol{\lambda}_k^{(\text{ii})}$ is obtained after randomly flipping the elements of existing binary vector $\boldsymbol{\lambda}_k$ based on a probability p_s. We chose either $\boldsymbol{\lambda}_k^{(\text{i})}$ or $\boldsymbol{\lambda}_k^{(\text{ii})}$ based on a simulated annealing

Line	Variable	Description
8	$u_1, .., u_D$	D independent random variables following a uniform distribution between 0 and 1.
9	p_s	A threshold used to discretize $u_1, .., u_D$
9	$\boldsymbol{f}$	A D dimensional vector obtained after discretizing $u_1, .., u_D$ based on p_s. It is used to flip bits in $\boldsymbol{\lambda}_k$ by XOR operation.
10	$\boldsymbol{\lambda}_k^{(\mathrm{i})}$, $\boldsymbol{\lambda}_k^{(\mathrm{ii})}$	Two candidate binary feature set vectors, $\boldsymbol{\lambda}_k^{(\mathrm{i})}$ is same as current $\boldsymbol{\lambda}_k$, $\boldsymbol{\lambda}_k^{(\mathrm{i})}$ is obtained after flipping elements of $\boldsymbol{\lambda}_k$ based on vector $\boldsymbol{f}$
11	$\langle \boldsymbol{\theta}_{k,\mathrm{I}}^{(\mathrm{i})}, \boldsymbol{\theta}_{k,\mathrm{NI}}^{(\mathrm{i})} \rangle$; $\langle \boldsymbol{\theta}_{k,\mathrm{I}}^{(\mathrm{ii})}, \boldsymbol{\theta}_{k,\mathrm{NI}}^{(\mathrm{ii})} \rangle$	Class boundary vectors corresponding to candidate experts using $\boldsymbol{\lambda}_k^{(\mathrm{i})}$, $\boldsymbol{\lambda}_k^{(\mathrm{ii})}$.
12	$\mathcal{L}^{(\mathrm{i})}$	Likelihood computed using $\boldsymbol{\theta}_{k,\mathrm{I}}^{(\mathrm{i})}; \boldsymbol{\theta}_{k,\mathrm{NI}}^{(\mathrm{i})}; \boldsymbol{\lambda}_k^{(\mathrm{i})}$
13	$\mathcal{L}^{(\mathrm{ii})}$	Likelihood computed using $\boldsymbol{\theta}_{k,\mathrm{I}}^{(\mathrm{ii})}; \boldsymbol{\theta}_{k,\mathrm{NI}}^{(\mathrm{ii})}; \boldsymbol{\lambda}_k^{(\mathrm{ii})}$ (Note parameters for other experts $k' \neq k$ remain same for computing $\mathcal{L}^{(\mathrm{i})}, \mathcal{L}^{(\mathrm{ii})}$ in iteration determining parameters for expert k.)
14	updated $\mathcal{L}^{(\mathrm{i})}$,$\mathcal{L}^{(\mathrm{ii})}$	New values of $\mathcal{L}^{(\mathrm{i})}$,$\mathcal{L}^{(\mathrm{i})}$ computed using updated $\langle \boldsymbol{\theta}_{k,\mathrm{I}}^{(\mathrm{i})}, \boldsymbol{\theta}_{k,\mathrm{NI}}^{(\mathrm{i})} \rangle$;$\boldsymbol{\lambda}_k^{(\mathrm{i})}$ and updated $\langle \boldsymbol{\theta}_{k,\mathrm{I}}^{(\mathrm{ii})}, \boldsymbol{\theta}_{k,\mathrm{NI}}^{(\mathrm{ii})} \rangle$;$\boldsymbol{\lambda}_k^{(\mathrm{ii})}$ after gradient ascent.

Table 2. Description of intermediate variables in Algorithm 1.

procedure. We perform gradient ascent on $\langle \boldsymbol{\theta}_{k,\mathrm{I}}; \boldsymbol{\theta}_{k,\mathrm{NI}}; \boldsymbol{\phi}_k \rangle$ on the two candidate feature subsets and retain the one which provides a higher ascent on the likelihood function. Algorithm 1 describes the modified algorithm and Table 2 explains the intermediate variables used in Algorithm 1.

4. EXPERIMENTS AND RESULTS

We perform a 5-fold cross validation using data from 44 speakers for training and 11 speakers for testing. The parameters $\langle K, p_s \rangle$ are tuned using an inner 2-fold cross validation on the training set. We use a single expert as our baseline. Table 3 reports the baseline accuracy, accuracy using vanilla MoE and MoE trained using the modified EM algorithm.

We observe a significant gain using an MoE (binomial proportions test, p value < 0.5) over a single expert. This supports our hypothesis that the variability in pathological data can be better modeled using an MoE. MoE based on the modified EM algorithm provides the best results suggesting that selecting a feature subset per expert provides a better intelligibility inference. In next section, we analyze the speaker assignment to experts and features selected for each expert.

5. DATA ANALYSIS

5.1. Distribution of speakers per expert

Each expert in an MoE learns a different class boundary based on the responsibilities R_{kn}. In this section, we investigate the patterns in modeling data samples from a single voice source. We study the distribution of data points from a single speaker amongst the experts based on the gating function. For this purpose, we train a model on entire dataset (55 speakers) and determine K based on the Akaike information criteria [27]. We obtain an MoE with $K = 3$ and empirically set p_s to 0.01.

We assemble all the data points from a given speaker S and compute the average gating function value $D(k, S)$ for expert k as shown in (7). A high value for $D(k, S)$ implies that the expert k is assigned high weight for data samples from speaker S. Note that $D(k, S)$

Algorithm 1 Modified EM algorithm for training MoE using simulated annealing-gradient ascent in the M-step.

Initialize: $\boldsymbol{\theta}_{k,\mathrm{I}}; \boldsymbol{\theta}_{k,\mathrm{NI}}; \boldsymbol{\phi}_k \; \forall k \in 1, .., K$
while Change in $\mathcal{L}$ is above a threshold **do**
  E-step:
  Compute R_{kn} for all combinations of k, n
    $k \in \{1, .., K\}$, $n \in \{1, .., N\}$
  M-step:
  for $k = 1$ to K **do**
    Sample $u_1, .., u_D$
    $\boldsymbol{f} = \{(u_1 < p_s), .., (u_D < p_s)\}$
    $\boldsymbol{\lambda}_k^{(\mathrm{i})} = \boldsymbol{\lambda}_k$; $\boldsymbol{\lambda}_k^{(\mathrm{ii})} = \mathrm{XOR}\ (\boldsymbol{\lambda}_k, \boldsymbol{f})$
    $\boldsymbol{\theta}_{k,\mathrm{I}}^{(\mathrm{i})} = \boldsymbol{\theta}_{k,\mathrm{I}}^{(\mathrm{ii})} = \boldsymbol{\theta}_{k,\mathrm{I}}$; $\boldsymbol{\theta}_{k,\mathrm{NI}}^{(\mathrm{i})} = \boldsymbol{\theta}_{k,\mathrm{NI}}^{(\mathrm{ii})} = \boldsymbol{\theta}_{k,\mathrm{NI}}$
    Gradient ascent (i): $\boldsymbol{\theta}_{k,\mathrm{I}}^{(\mathrm{i})}, \boldsymbol{\theta}_{k,\mathrm{NI}}^{(\mathrm{i})}, \boldsymbol{\phi}_k^{(\mathrm{i})}$ on $\mathcal{L}^{(\mathrm{i})}$
    Gradient ascent (ii): $\boldsymbol{\theta}_{k,\mathrm{I}}^{(\mathrm{ii})}, \boldsymbol{\theta}_{k,\mathrm{NI}}^{(\mathrm{ii})}, \boldsymbol{\phi}_k^{(\mathrm{ii})}$ on $\mathcal{L}^{(\mathrm{ii})}$
    if (updated $\mathcal{L}^{(\mathrm{i})}$) $>$ (updated $\mathcal{L}^{(\mathrm{ii})}$) **then**
      Update:$\langle \boldsymbol{\theta}_{k,\mathrm{I}}; \boldsymbol{\theta}_{k,\mathrm{NI}}; \boldsymbol{\phi}_k; \mathcal{L} \rangle$=$\langle \boldsymbol{\theta}_{k,\mathrm{I}}^{(\mathrm{i})}, \boldsymbol{\theta}_{k,\mathrm{NI}}^{(\mathrm{i})}; \boldsymbol{\phi}_k^{(\mathrm{i})}; \mathcal{L}^{(\mathrm{i})} \rangle$
    else
      Update:$\langle \boldsymbol{\theta}_{k,\mathrm{I}}; \boldsymbol{\theta}_{k,\mathrm{NI}}; \boldsymbol{\phi}_k; \mathcal{L} \rangle$=$\langle \boldsymbol{\theta}_{k,\mathrm{I}}^{(\mathrm{ii})}, \boldsymbol{\theta}_{k,\mathrm{NI}}^{(\mathrm{ii})}; \boldsymbol{\phi}_k^{(\mathrm{ii})}; \mathcal{L}^{(\mathrm{ii})} \rangle$
    end if
  end for
end while

Classifier	Accuracy	Class recall	
		NI	I
Baseline	62.88	61.98	63.77
Vanilla MoE	65.07	63.93	66.19
MoE using modified EM	**66.00**	64.86	67.11

Table 3. Classification results

is a probability mass function over $k = \{1..K\}$ as the values are non-negative and sum to one. Figure 1 shows the values for $D(k, S)$ $(k = 1, 2, 3)$ for all the 55 speakers.

$$D(k, S) = \frac{\sum\limits_{\boldsymbol{x}_n \text{ from speaker } S} p(m_k | \boldsymbol{x}_n)}{\text{Number of data points from } S} \tag{7}$$

From the figure, we observe an uneven distribution for $D(k, S)$ across different speakers. This implies that experts are weighted unequally for a given speaker. 30 speakers have a $D(k, S)$ value higher than 0.5 for a single expert. We compare the entropy of obtained $D(k, S)$ against a strategy giving equal weight to each expert for every speaker ($D(k, S) \sim$ Uniform distribution). We compute the entropy $E(S)$ of the distribution $D(k, S)$ for every speaker as shown in (8). We then compare the mean of obtained entropies against entropy of a uniform distribution. The t-statistic (t) for a one sample t-test as shown in (9). We observe a significant difference (p-value $<$ .001) suggesting a non-uniform allocation of experts on data samples per speaker. Figure 2 shows plot of two arbitrarily chosen features from speaker 51 and 55 with highest $D(k, S)$ values for $k = 2$ and $k = 1$, respectively. We see that the data samples from these two speakers appear as separate clusters and note a difference in class boundaries from the experts over the plotted features. These clusters in feature space may form due to several reasons as speaker traits as well as similarity of in pathological conditions amongst patients. An MoE can model such data clusters better than a single expert, as is also suggested by our results.

$$E(S) = -\sum_{k=1}^{K} D(k, S) \log D(k, S) \tag{8}$$

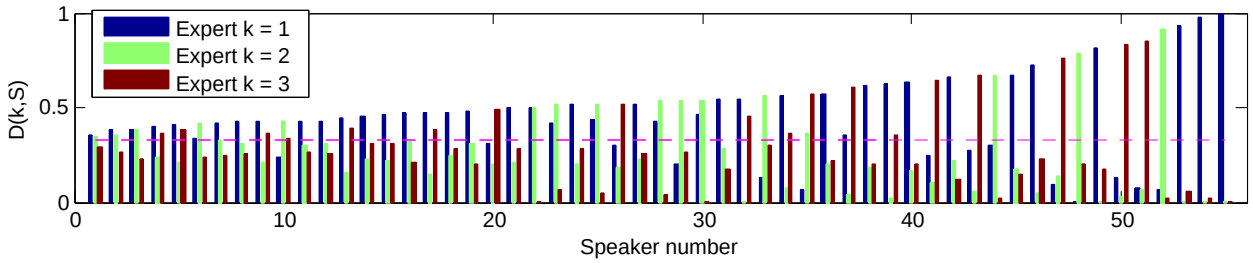

Fig. 1. Distribution of expert assignment over data points from a given speaker. We sort the speakers based on the entropy $E(S)$. The magenta line shows $D(k, S)$ level for a uniform distribution.

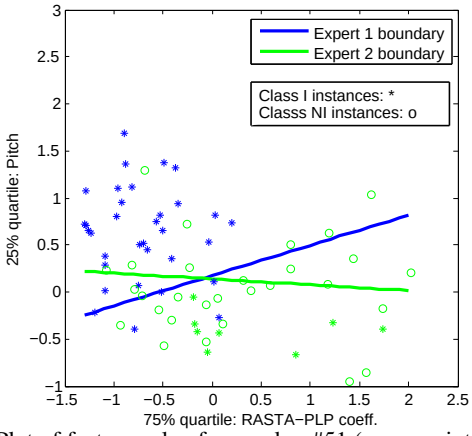

Fig. 2. Plot of feature value for speaker #51 (green points) and #55 (blue points).

$$t = \frac{\text{Mean of } E(S) - \text{Entropy for uniform distribution}}{\text{Std. deviation of } E(S)/\sqrt{\text{No. of speakers}}} \quad (9)$$

We further investigate the relation between how well a speaker is modeled and the entropy of the distribution $D(k, S)$. For a given speaker, we define the average log-likelihood over his data samples ($\mathcal{L}(S)$, see (10)) as the goodness measure. We fit a linear model to predict $\mathcal{L}(S)$ using $E(S)$ as shows in (11). A lower entropy implies more biased allocation of speaker's data samples. $\langle a; b\rangle$ are parameters determined using linear regression. The fit and statistics are shown in Figure 3.

$$\mathcal{L}(S) = \frac{\sum_{\boldsymbol{x}_n \text{ from speaker} S} \log p(y_n|\boldsymbol{x}_n)}{\text{Number of data points from } S} \quad (10)$$

$$\mathcal{L}(S) = a \times E(S) + b \quad (11)$$

A negative value for a shows that $\mathcal{L}(S)$ increases as the entropy decreases. F-test on linear regression rejects the null hypothesis of a constant model at 5% significance level. This provides some evidence that a higher consistency in expert selection leads to better modeling. Hence not only we observe a biased modeling of a speaker's data points, s/he is modeled better as the bias increases. This biased assignment of experts encourages us to investigate any relation between the speakers' characteristics and expert weighting. As each expert is assigned high weights for a few speakers, any relation between experts, corresponding speakers and their diagnosis can be investigated .

5.2. Feature subsets per expert

We observe that our experts retain a majority of the features per expert. For the MoE trained over all the data (in section 5.1), the three experts retain 230, 232 and 235 features and rejected features are mutually exclusive across experts. Although simulated annealing provides a suboptimal greedy solution, our experiment suggests features rejected by one expert are used by another. This shows that the feature utility varies across the feature space. The feature utility can be subject to further investigation based on the nature of pathological diagnosis.

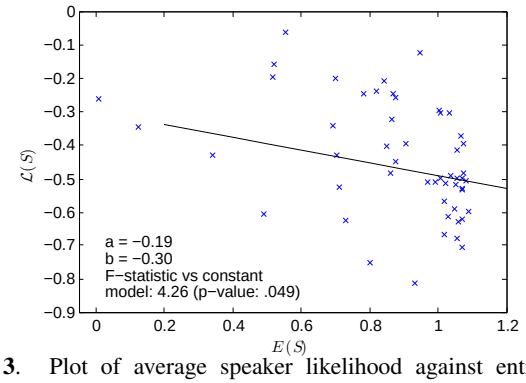

Fig. 3. Plot of average speaker likelihood against entropy of $D(k, S)$. A linear fit (black line) suggests an increase in likelihood with a more biased distribution.

6. CONCLUSION

Pathological speech may be caused by several factors and intelligibility analysis may inform us regarding the pathological conditions. However this is a challenging problem given the variety of reasons leading to a compromised voice production. We apply an MoE framework to address the complex learning problem as well as providing interpretability to aid diagnosis. Furthermore, we modify the MoE learning algorithm to incorporate joint feature selection for each expert. We show that MoE trained using the modified algorithm performs better than a single expert and a vanilla MoE. We observe a biased allocation of a speaker's data samples to experts in the mixture. A linear fit shows that a higher consistency in expert assignment leads to higher data log-likelihood for a given speaker. These experiments reflect that certain experts provide better expertise for a given speaker and provide motivation for analysis based on pathological diagnosis.

Our experiments encourage further investigation with availability of detailed pathological diagnosis. We aim to analyze the relation of each expert and the pathological condition. Furthermore, the feature subset for each expert may inform us regarding the nature of pathological conditions and the aspects of voice affected. The modeling approach may be further refined based on experimentation with other expert models and binary integer programming methods. The modified EM algorithm provides a generic tool for data analysis and can be applied to other domains characterized by similar data characteristics.

7. REFERENCES

[1] S Hadjitodorov and P Mitev, "A computer system for acoustic analysis of pathological voices and laryngeal diseases screening," *Medical engineering & physics*, vol. 24, no. 6, pp. 419–429, 2002.

[2] R Jones, "Observations on stammering after localized cerebral injury.," *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 29, no. 3, pp. 192, 1966.

[3] C Lai, S Fisher, J Hurst, F Vargha-Khadem, and A Monaco, "A forkhead-domain gene is mutated in a severe speech and language disorder," *Nature*, vol. 413, no. 6855, pp. 519–523, 2001.

[4] C Van Riper and R Erickson, *Speech correction: An introduction to speech pathology and audiology*, Allyn and Bacon, 1996.

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

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

[7] A Dibazar, T Berger, and S Narayanan, "Pathological voice assessment," in *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*. IEEE, 2006, pp. 1669–1673.

[8] D Michaelis, M Fröhlich, and H Strube, "Selection and combination of acoustic features for the description of pathologic voices," *The Journal of the Acoustical Society of America*, vol. 103, no. 3, pp. 1628–1639, 1998.

[9] A Dibazar, S Narayanan, and T Berger, "Feature analysis for automatic detection of pathological speech," in *Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint*. IEEE, 2002, vol. 1, pp. 182–183.

[10] J Godino-Llorente, P Gomez-Vilda, and M Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters," *Biomedical Engineering, IEEE Transactions on*, vol. 53, no. 10, pp. 1943–1953, 2006.

[11] R Ritchings, M McGillion, and C Moore, "Pathological voice quality assessment using artificial neural networks," *Medical engineering & physics*, vol. 24, no. 7, pp. 561–564, 2002.

[12] R Behroozmand and F Almasganj, "Optimal selection of wavelet-packet-based features using genetic algorithm in pathological assessment of patients speech signal with unilateral vocal fold paralysis," *Computers in Biology and Medicine*, vol. 37, no. 4, pp. 474–485, 2007.

[13] M Arjmandi and M Pooyan, "An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine," *Biomedical Signal Processing and Control*, vol. 7, no. 1, pp. 3–19, 2012.

[14] C Middag, J Martens, G Van Nuffelen, and M De Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 3, 2009.

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

[16] J Kim, N Kumar, A Tsiartas, and M Li, "Automatic intelligibility classification of sentence-level pathological speech," *Computer, Speech, and Language*, 2014.

[17] D Huang, Y Zhu, D Wu, and R Yu, "Detecting intelligibility by linear dimensionality reduction and normalized voice quality hierarchical features.," in *INTERSPEECH*, 2012.

[18] D Huang, M Dong, and H Li, "Intelligibility detection of pathological speech using asymmetric sparse kernel partial least squares classifier," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3744–3748.

[19] B Peralta and A Soto, "Embedded local feature selection within mixture of experts," *Information Sciences*, vol. 269, pp. 176–187, 2014.

[20] A Khalili, "New estimation and feature selection methods in mixture-of-experts models," *Canadian Journal of Statistics*, vol. 38, no. 4, pp. 519–539, 2010.

[21] B Peralta, "Simultaneous feature and expert selection within mixture of experts," *arXiv preprint arXiv:1405.7624*, 2014.

[22] R Gupta, K Audhkhasi, and S Narayanan, "Training ensemble of diverse classifiers on feature subsets," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing (ICASSP)*, May 2014.

[23] L van der Molen, M van Rossum, A Ackerstaff, L Smeele, C Rasch, and F Hilgers, "Pretreatment organ function in patients with advanced head and neck cancer: clinical outcome measures and patients' views," *BMC Ear, Nose and Throat Disorders*, vol. 9, no. 1, pp. 10, 2009.

[24] F Eyben, M Wöllmer, and B Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[25] Kenji Kira and Larry A Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *AAAI*, 1992, pp. 129–134.

[26] P Moerland, "Some methods for training mixtures of experts," *IDIAP Communication, Dalle Molle Institute for Perceptive Artificial Intelligence*, 1997.

[27] K Yamaoka, T Nakagawa, and T Uno, "Application of akaike's information criterion (aic) in the evaluation of linear pharmacokinetic equations," *Journal of pharmacokinetics and biopharmaceutics*, vol. 6, no. 2, pp. 165–175, 1978.