

Normalizing Flow based Hidden Markov Models for Classification of Speech Phones with Explainability

Anubhab Ghosh, Antoine Honoré, Dong Liu, Gustav Eje Henter, Saikat Chatterjee

Digital Futures, and School of Electrical Engg. and Computer Sc., KTH Royal Institute of Technology, Sweden

anubhabg@kth.se, honore@kth.se, doli@kth.se, ghe@kth.se, sach@kth.se

Abstract—In pursuit of explainability, we develop generative models for sequential data. The proposed models provide state-of-the-art classification results and robust performance for speech phone classification. We combine modern neural networks (normalizing flows) and traditional generative models (hidden Markov models - HMMs). Normalizing flow-based mixture models (NMMs) are used to model the conditional probability distribution given the hidden state in the HMMs. Model parameters are learned through judicious combinations of time-tested Bayesian learning methods and contemporary neural network learning methods. We mainly combine expectation-maximization (EM) and mini-batch gradient descent. The proposed generative models can compute likelihood of a data and hence directly suitable for maximum-likelihood (ML) classification approach. Due to structural flexibility of HMMs, we can use different normalizing flow models. This leads to different types of HMMs providing diversity in data modeling capacity. The diversity provides an opportunity for easy decision fusion from different models. For a standard speech phone classification setup involving 39 phones (classes) and the TIMIT dataset, we show that the use of standard features called mel-frequency-cepstral-coefficients (MFCCs), the proposed generative models, and the decision fusion together can achieve 86.6% accuracy by generative training only. This result is close to state-of-the-art results, for examples, 86.2% accuracy of PyTorch-Kaldi toolkit [1], and 85.1% accuracy using light gated recurrent units [2]. We do not use any discriminative learning approach and related sophisticated features in this article.

Keywords: Phone recognition, generative models, hidden Markov models, neural networks.

I. INTRODUCTION

Neural networks based discriminative methods and generative models based maximum-likelihood (ML) methods are two main directions in pattern classification. ML classification is an optimal rule derived from Bayes minimum risk criterion under certain technical conditions.

Neural networks and their deep versions in discriminative setups are successful to provide state-of-the-art classification performances across many applications. Neural networks are data-driven, model-free, and are typically optimized for a pre-defined number of classes. In many cases, structures of neural network based deep systems are not self-explanatory. It is difficult to understand individual roles of components or learning tricks in a deep system. This limitation in understanding is a major reason for lack of explainability.

On the other hand, model-based systems can be subjected to scrutiny and analysis for understanding. Generative models are typically model-based and can explain the process of data generation. If generative models can compute likelihood of a

data point then they are suitable for ML classification. For example, a Gaussian mixture model (GMM) is a suitable generative model widely used for ML classification. Parameters of a GMM are learned from data using time-tested learning principles, such as expectation-maximization (EM). GMMs have been used in numerous applications with robust performance where data is corrupted. Our opinion is that generative models with explainable data generation process, use of time-tested learning principles, and scope of robust performance for many potential applications provide a path towards explainable machine learning and trust.

In pursuit of explainability, we combine advantages of model-free neural networks and generative model-based ML classification systems. For modeling sequential data, we use hidden Markov models (HMMs) as a basic scheme. Then we develop new generative models by combining neural networks and HMMs. Performances of the proposed models are demonstrated for a speech recognition application in this article. While HMMs have been widely used for speech recognition [3], they have been used in many other applications, such as handwriting recognition [4], activity recognition [5], genomic signal processing [6], transport forecasting [7], etc.

For speech recognition, the conditional probability distribution of a state in HMM is classically modelled using GMM. The GMM-HMM combination is a pure model-based system. GMM-HMMs are generative models that can be trained using time-tested probabilistic machine learning methods, such as EM, variational Bayes (VB) and Markov-chain-monte-carlo (MCMC). In spite of success, GMM-HMMs have limited modeling capability owing to the use of GMMs for representing complicated data manifolds.

We improve modeling capability of the conditional state distributions in an HMM. We use neural networks for the conditional state distributions. There are recent advances in probability density modeling using neural networks called normalizing flows [8]. Normalizing-flows are generative models where we can compute likelihood of a data point. Later, mixtures of normalizing flows were proposed and EM based training was shown in [9]. A mixture of normalizing flows is a generative model that can handle multiple modes and manifolds of a data distribution. We refer to the proposed models as normalizing flow-based mixture models (NMMs). For our proposed models, we use NMMs as the conditional distributions of HMM states. That means, we replace time-tested GMMs by modern NMMs. The combination of NMM and HMM is referred to as NMM-HMM in contrast to GMM-

HMM [10].

There are several varieties of normalizing flows. We explore the use of RealNVP [11] and Glow [12] models in our work, and we use succinct notations such as NVP-HMM and Glow-HMM to describe corresponding NMM variety in our work. There are other normalizing flow models, for example, autoregressive (AR) flow [13], that also fit in our development. We do not explore the other flow models due to high complexity.

This article builds on the prior works [9], [10], [14]. Our main technical contributions in this article are as follows.

- We develop NMM-HMM models (NVP-HMM and Glow-HMM) and show their better performance than GMM-HMM for a robust speech recognition problem under ML classification principle. We do not use any discriminative learning approach.
- We formulate judicious combinations of EM and gradient search (back-propagation) that can learn the parameters of NMM-HMMs.
- We show that appropriate combination of GMM-HMM and proposed NMM-HMMs using a simple decision fusion strategy yields state-of-the-art results. The decision fusion also provides robust classification performance. Our results conclusively shows generative model based ML classification is competitive.

We performed extensive experiments for phone recognition using implementations based on Python, Kaldi and PyTorch. We used the TIMIT database and 39 phone classes as per standard practice in literature [15], [16].

A. Relevant literature

There is a growing literature on explainable machine learning, with a significant interest on (deep) neural networks. Existing methods for explanations can be categorized in various ways, [17], for example, *intrinsic* and *post-hoc* [18]. Intrinsic explainability is inherent in structures of methods, such as rule-based models, decision trees, linear models, and unfolding neural networks [19]. In contrast, a post-hoc method requires a second (simple) model to provide explanations about the original (complex) method. Therefore the proposed generative models in this article are close to intrinsic nature of explainability. There are other kind of approaches to provide *local* and *global* explanations. A local explanation justifies a model or method’s output for a specific input. A global explanation provides a justification on average performance of a model, independently of any particular input [20]. There are techniques for explainability using visualization of neuronal activity in layers of a deep architecture, for examples, based on sparsity and heatmaps [21], [22], [23]. Another approach of explainability depends on explaining contributions of input features to output, referred to as feature importance [24]. In the background of this literature survey, our proposed models combine model-based and data-driven methods for explaining data generation process. We also perform robustness study of our proposed models using corrupted data at varying noise conditions. Overall, our proposed models have connections with intrinsic explainability and global explainability.

We now provide a literature survey on speech phone classification. For the TIMIT dataset, a list of phone recognition accuracy results on 39 phones is summarized in Github [25]. According to the list, the best phone recognition accuracy is 86.2% [1]. The work in [2] based on light gated recurrent units reported 85.1% accuracy. The work [26] based on hierarchical convolutional deep maxout networks reported 83.5% accuracy. Therefore, the ballpark accuracy for TIMIT phone recognition today is around 85 – 86%.

Prior to state-of-the-art results, an earlier attempt was to use restricted Boltzmann machines (RBMs) to form a deep belief network (DBN) that served as the acoustic models for HMMs [27]. The DBN used a softmax output layer and was discriminatively trained using backpropagation, achieving a phone recognition accuracy around 77% on the TIMIT dataset.

Dynamical neural networks have also been used for speech recognition. Examples are recurrent neural networks (RNNs), long-short-term-memory networks (LSTMs) [28], and their gated recurrent unit based modifications. Attention mechanisms are also found to provide 82.4% accuracy [29]. Almost all these example works employ discriminative learning.

There are end-to-end designs to achieve good performance for speech recognition [29], [30]. These methods typically learn appropriate features. In this article, we deliberately avoid feature learning or use of sophisticated features [31], [32], [33]. The reason is that we wish to test the power of proposed generative models solely. We perform all our experiments using time-tested features called MFCCs [34]. Development of MFCCs uses traditional speech signal processing knowledge.

II. NMM-HMM

Let the number of classes be C for ML classification. We denote the generative model for the c ’th class as \mathbf{H}_c . We denote a data sequence $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T]^\top$ (\top denotes the transpose operator), where $\mathbf{x}_t \in \mathbb{R}^D$ denotes the feature vector at time t , and T denotes the sequence length. Then the time-tested ML-classification approach is

$$\mathbf{H}_c^* = \arg \max_{H_c} p(\mathbf{x}|\mathbf{H}_c), \quad (1)$$

where $p(\mathbf{x}|\mathbf{H}_c)$ denotes likelihood of the data sequence \mathbf{x} given the model \mathbf{H}_c . Here \mathbf{H}_c can be a GMM-HMM or a NMM-HMM. We explain NMM-HMM below.

A. NMM-HMM architecture

We now drop the subscript c for notational clarity. An HMM is characterized by the following quantities: the set of hidden states of the HMM denoted by S , the initial probability vector for the states of the HMM is denoted by \mathbf{q} , the state-transition probability matrix is denoted by \mathbf{A} , and the conditional probability distribution given the hidden state s is denoted by $p(\mathbf{x}|s; \Psi_s)$, where Ψ_s are the parameters. The conditional distribution given the hidden state (s) in NMM-HMM is modelled as a weighted mixture of N_{mix} density functions as

$$p(\mathbf{x}|s; \Psi_s) = \sum_{k=1}^{N_{mix}} \pi_{s,k} p(\mathbf{x}|s; \Phi_{s,k}). \quad (2)$$

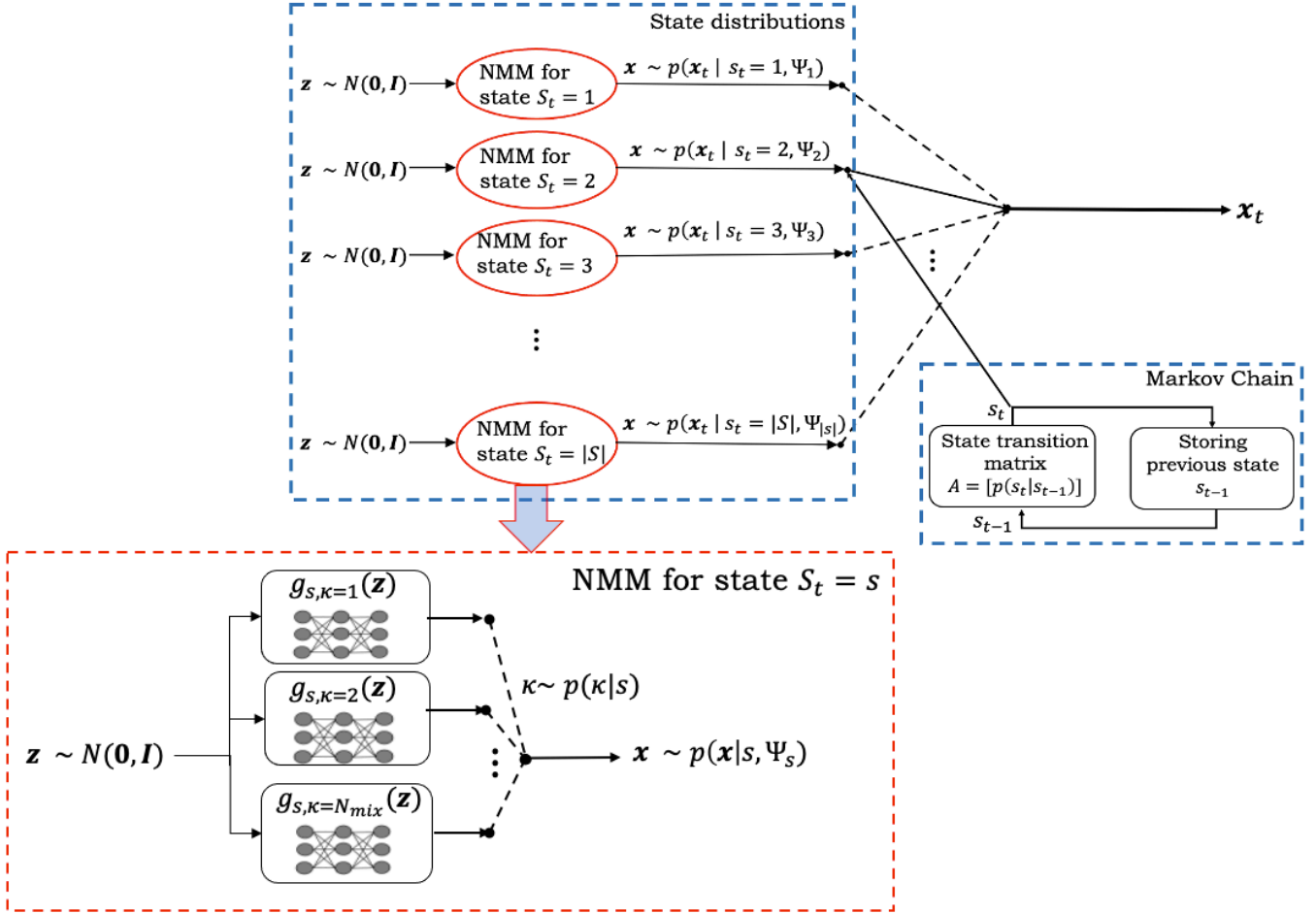


Fig. 1: Schematic of the NMM-HMM architecture. The architecture inside a single NMM model for a given state $S_t = s$ is also shown in an expanded view at the bottom of the figure using red dotted line.

In (2), the weights $\{\pi_{s,k}\}$ denote the probability of drawing a given mixture component k from a categorical distribution with $\pi_{s,k} = p(k|s; \mathbf{H})$ and they satisfy $\sum_{k=1}^{N_{mix}} \pi_{s,k} = 1$. We use normalizing-flow based density function as the mixture component, denoted by $p(\mathbf{x}|s; \Phi_{s,k})$. A normalizing-flow function is a neural network $\mathbf{g}_{s,k} : \mathbb{R}^D \rightarrow \mathbb{R}^D$, such that

$$\mathbf{x} = \mathbf{g}_{s,k}(\mathbf{z}),$$

where \mathbf{z} is a D -dimensional latent variable that is drawn from a known prior distribution. We assume $p(\mathbf{z})$ is a *standard* isotropic Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The function $\mathbf{g}_{s,k}$ is invertible. The inverse function $\mathbf{f}_{s,k}$ (or equivalently $\mathbf{g}_{s,k}^{-1}$) is also called *normalizing* function. The use of the term ‘normalizing’ is due to the reason that the signal \mathbf{x} is generated from \mathbf{z} sampled from a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ [8]. The inverse function is mathematically described as

$$\mathbf{z} = \mathbf{f}_{s,k}(\mathbf{x}) \triangleq \mathbf{g}_{s,k}^{-1}(\mathbf{x}).$$

Using the change of variable formula in probability theory, we can calculate the likelihood for every mixture component as

$$p(\mathbf{x}|s; \Phi_{s,k}) = p(\mathbf{f}_{s,k}(\mathbf{x}) | \det(\mathbf{J}(\mathbf{f}_{s,k}(\mathbf{x})))|, \quad (3)$$

where $\mathbf{J}(\mathbf{f}_{s,k}(\mathbf{x})) = \left[\frac{\partial \mathbf{f}_{s,k}(\mathbf{x})}{\partial \mathbf{x}} \right]$, denoting the Jacobian of the normalizing function $\mathbf{f}_{s,k}(\mathbf{x})$. The mixture model in (2) using normalizing flows, with each mixture component defined using (3), is referred to as Normalizing flow-based mixture model (NMM). The NMM model is

$$\begin{aligned} p(\mathbf{x}|s; \Psi_s) &= \sum_{k=1}^{N_{mix}} \pi_{s,k} p(\mathbf{f}_{s,k}(\mathbf{x}) | \det(\mathbf{J}(\mathbf{f}_{s,k}(\mathbf{x})))| \\ &= \sum_{k=1}^{N_{mix}} \pi_{s,k} p(\mathbf{g}_{s,k}^{-1}(\mathbf{x}) | \det(\mathbf{J}(\mathbf{g}_{s,k}^{-1}(\mathbf{x})))|. \end{aligned} \quad (4)$$

The schematic of the NMM-HMM architecture is shown in Fig. 1.

It should be noted that NMMs satisfy certain specific properties in order to be incorporated into an existing probabilistic framework such as hidden Markov models. One important property is that the mapping defined by each mixture component from \mathbf{z} to \mathbf{x} is bijective. Thus, the mapping is one-to-one and an inverse mapping from \mathbf{x} to \mathbf{z} also exist. Also, the fact that the likelihood values can be computed exactly for an NMM enable similar mathematical treatment as a Gaussian mixture model (GMM). The log-determinant of the Jacobian

$\mathbf{J}(\mathbf{f}_{s,k}(\mathbf{x}))$ should be also easy to compute. Each mixture component of an NMM for a given state has its own set of parameters $\Phi_{s,k}$, which require to be learned using back-propagation. This is illustrated by explanation of the learning problem and its solution in the subsequent section.

B. Learning of NMM-HMM parameters

Let us assume we have M number of training data sequences as $\{\mathbf{x}^{(m)}\}_{m=1}^M$. Then the ML-based learning problem is:

$$\begin{aligned} \mathbf{H}^* &= \arg \max_{\mathbf{H} \in \mathcal{H}} \frac{1}{M} \prod_{m=1}^M p(\mathbf{x}^{(m)}; \mathbf{H}) \\ &\triangleq \arg \max_{\mathbf{H} \in \mathcal{H}} \frac{1}{M} \log \left(\prod_{m=1}^M p(\mathbf{x}^{(m)}; \mathbf{H}) \right) \\ &= \arg \max_{\mathbf{H} \in \mathcal{H}} \frac{1}{M} \sum_{m=1}^M \log \left(p(\mathbf{x}^{(m)}; \mathbf{H}) \right), \end{aligned} \quad (5)$$

where the possible hypothesis set of models is denoted by \mathcal{H} . The learning problem in (5) can be solved using the well known expectation-maximization (EM) framework [35]. The expectation step ("E-step") involves the calculation of the posterior probability distribution of hidden sequences of states \mathbf{s} and mixture components \mathbf{k} , to obtain an expected value of the log-likelihood of the data sequence \mathbf{x} under the current model parameters \mathbf{H}^{old} . This is formulated as a cost function $\mathcal{L}(\mathbf{H}; \mathbf{H}^{old})$ and is shown in (6) as

$$\mathcal{L}(\mathbf{H}; \mathbf{H}^{old}) = \mathbf{E}_{p(\mathbf{s}, \mathbf{k} | \mathbf{x}; \mathbf{H}^{old})} \log(p(\mathbf{x}, \mathbf{s}, \mathbf{k}; \mathbf{H})). \quad (6)$$

The second step, known as the maximisation step ("M-step"), consists of finding the model \mathbf{H}^* that maximizes the expected log-likelihood computed in (6). This can be decomposed into three separate maximization problems as

$$\begin{aligned} &\max_{\mathbf{H}} \mathcal{L}(\mathbf{H}; \mathbf{H}^{old}) \\ &= \max_{\mathbf{q}} \mathcal{L}(\mathbf{q}; \mathbf{H}^{old}) + \max_{\mathbf{A}} \mathcal{L}(\mathbf{A}; \mathbf{H}^{old}) + \max_{\Psi} \mathcal{L}(\Psi; \mathbf{H}^{old}), \end{aligned} \quad (7)$$

where,

$$\mathcal{L}(\mathbf{q}; \mathbf{H}^{old}) = \mathbf{E}_{p(\mathbf{s} | \mathbf{x}; \mathbf{H}^{old})} \log(p(s_1; \mathbf{H})), \quad (8)$$

$$\mathcal{L}(\mathbf{A}; \mathbf{H}^{old}) = \mathbf{E}_{p(\mathbf{s} | \mathbf{x}; \mathbf{H}^{old})} \sum_{t=2}^T \log(p(s_t | s_{t-1}; \mathbf{H})), \quad (9)$$

$$\mathcal{L}(\Psi; \mathbf{H}^{old}) = \mathbf{E}_{p(\mathbf{s}, \mathbf{k} | \mathbf{x}; \mathbf{H}^{old})} \log(p(\mathbf{x}, \mathbf{k} | \mathbf{s}; \mathbf{H})). \quad (10)$$

The maximisation problems described in (8), (9) can be solved using standard EM forward-backward algorithm to compute the posterior distribution effectively, explained in [35]. For solving the maximisation problem in (10), we need to maximise the log-likelihood with respect to the mixture of weights and the set of parameters of the flow models. This would require the computation of the log-determinant of the appropriate normalizing flow model. For solution of (10), the problem can be broken down into two sub-problems, i.e. learning the set of mixture of weights $\mathbf{\Pi} = \{\pi_s | s \in S\}$, and the

learning the set of flow model parameters $\Phi = \{\phi_s | s \in S\}$. This is shown as

$$\max_{\Psi} \mathcal{L}(\Psi; \mathbf{H}^{old}) = \max_{\mathbf{\Pi}} \mathcal{L}(\mathbf{\Pi}; \mathbf{H}^{old}) + \max_{\Phi} \mathcal{L}(\Phi; \mathbf{H}^{old}). \quad (11)$$

The problem of learning the mixture of weights can be solved using a simple Lagrangian formulation while problem of learning the flow model parameters can be solved by using the results of the change of variable formula derived in (3) as

$$\begin{aligned} \mathcal{L}(\Phi; \mathbf{H}^{old}) &= \mathbf{E}_{p(\mathbf{s}, \mathbf{k} | \mathbf{x}; \mathbf{H}^{old})} \log(p(\mathbf{x} | \mathbf{s}, \mathbf{k}; \mathbf{H})) \\ &= \mathbf{E}_{p(\mathbf{s}, \mathbf{k} | \mathbf{x}; \mathbf{H}^{old})} [\log(p(\mathbf{f}_{s,k}(\mathbf{x}))) \\ &\quad + \log(|\det(\nabla \mathbf{f}_{s,k})|)], \end{aligned} \quad (12)$$

The first term on the right hand side of (12) is an expectation computed over the log-probability of latent data derived using the inverse function $\mathbf{f}_{s,k}$, and the second term is the result of the log-determinant of the Jacobian that is computed for the type of normalizing flow used. In the subsequent section, we explain the flow models that we used for modeling, and the computation of the log-determinant of the Jacobian ($\log(|\det(\nabla \mathbf{f}_{s,k})|)$).

C. RealNVP flow

We refer to a normalizing flow-based HMM (NMM-HMM) using RealNVP flow as NVP-HMM. The signal flow for a flow model having L layers can be illustrated as

$$\mathbf{z} = \mathbf{h}_0 \xrightarrow[\mathbf{f}_{s,k}^{[1]}]{\mathbf{g}_{s,k}^{[1]}} \mathbf{h}_1 \xrightarrow[\mathbf{f}_{s,k}^{[2]}]{\mathbf{g}_{s,k}^{[2]}} \mathbf{h}_2 \xrightarrow[\mathbf{f}_{s,k}^{[3]}]{\mathbf{g}_{s,k}^{[3]}} \mathbf{h}_3 \dots \xrightarrow[\mathbf{f}_{s,k}^{[L]}]{\mathbf{g}_{s,k}^{[L]}} \mathbf{h}_L = \mathbf{x}, \quad (13)$$

where $\mathbf{f}_{s,k}^{[l]}$ denotes the l^{th} layer of the network $\mathbf{f}_{s,k}$, and each such $\mathbf{f}_{s,k}^{[l]}$ is invertible. One of the first variety of normalizing flows have been proposed in [36], and different varieties discussed elaborately in [8]. Different flow model architectures may have different kinds of *coupling* between two successive layers in the network. An example can be obtained from the architecture called RealNVP flow that is discussed in [11]. To illustrate a small section of the mapping from the data space to the latent space, let us consider the input feature at the l^{th} layer denoted by \mathbf{h}_l . This feature is mapped to \mathbf{h}_{l-1} using the function $\mathbf{f}_{s,k}$. At every layer the D -dimensional input feature is split into two disjoint parts. Let us assume the features are $[\mathbf{h}_{l,1:d}, \mathbf{h}_{l,d+1:D}]^T$ (where d denotes the number of components in the first sub part). The relation is

$$\begin{aligned} \mathbf{h}_{l,1:d} &= \mathbf{h}_{l-1,1:d}, \\ \mathbf{h}_{l,d+1:D} &= \mathbf{h}_{l-1,d+1:D} \odot \exp(\mathbf{s}(\mathbf{h}_{l-1,1:d})) + \mathbf{t}(\mathbf{h}_{l-1,1:d}), \end{aligned} \quad (14)$$

where the symbol \odot denotes element-wise multiplications, $\mathbf{s} : \mathbb{R}^d \rightarrow \mathbb{R}^{D-d}$, $\mathbf{t} : \mathbb{R}^d \rightarrow \mathbb{R}^{D-d}$, with \mathbf{s}, \mathbf{t} being shallow feed-forward neural networks that differ only in the activation function for the last layer, which is a hyperbolic tangent (\tanh) activation function for modeling the logarithm of the standard deviation and an identity activation for the translation parameter [11]. The inverse mapping from the

data space to the latent space is defined as

$$\begin{aligned} \mathbf{h}_{l-1,1:d} &= \mathbf{h}_{l,1:d}, \\ \mathbf{h}_{l-1,d+1:D} &= (\mathbf{h}_{l,d+1:D} - \mathbf{t}(\mathbf{h}_{l,1:d})) \odot \exp(-\mathbf{s}(\mathbf{h}_{l,1:d})), \end{aligned} \quad (15)$$

A single layer of the mapping described in (14), (15), is referred to as a *coupling layer* [11]. For the flow model, the inverse mapping shown in (15) is computed for every layer, and the determinant of the Jacobian matrix is computed as the product of the layer-wise determinants of the Jacobian matrices as

$$\det(\nabla \mathbf{f}_{s,k}) = \prod_{l=1}^L \det(\nabla \mathbf{f}_{s,k}^{[l]}), \quad (16)$$

where each $\det(\nabla \mathbf{f}_{s,k}^{[l]})$ is computed as

$$\begin{aligned} \det(\nabla \mathbf{f}_{s,k}^{[l]}) &= \det\left(\begin{bmatrix} \mathbf{I}_{1:d} & \mathbf{0} \\ \frac{\partial \mathbf{h}_{l-1,d+1:D}}{\partial \mathbf{h}_{l-1,1:d}} & \text{diag}(-\mathbf{s}(\mathbf{h}_{l,1:d})) \end{bmatrix}\right) \\ &= \det(\text{diag}(-\mathbf{s}(\mathbf{h}_{l,1:d}))). \end{aligned} \quad (17)$$

In (17), $\mathbf{I}_{1:d}$ denotes an identity matrix and $\text{diag}(\dots)$ denotes a diagonal matrix with the elements of the argument vector in the main diagonal. It is important to ensure an alternate ordering between the two parts of the signal $[\mathbf{h}_{l,1:d}, \mathbf{h}_{l,d+1:D}]^T$ (described in the affine coupling layer) in (14) so that the required transformation is achieved. If the mapping in the l^{th} layer is defined in (14), then the mapping in the $(l+1)^{\text{th}}$ layer is defined as

$$\begin{aligned} \mathbf{h}_{l+1,1:d} &= \mathbf{h}_{l,1:d} \odot \exp(\mathbf{s}(\mathbf{h}_{l,d+1:D})) + \mathbf{t}(\mathbf{h}_{l,d+1:D}), \\ \mathbf{h}_{l+1,d+1:D} &= \mathbf{h}_{l,d+1:D}. \end{aligned} \quad (18)$$

D. Generative 1×1 convolution flow (Glow)

A variation of the RealNVP model used for generating image signals is known as Glow (Generative 1×1 convolution based flow) [12]. Similar to RealNVP, a normalizing flow-based HMM (NMM-HMM) using Glow is referred to as Glow-HMM. It introduces two important modifications: an *activation normalization* layer to standardise the activations at every flow step and an *invertible convolution* layer as a generalization of the switching operation between channels. In its simplest form, a Glow model consists of layers that are referred to as *flow steps* (analogous to L layers in case of RealNVP). The signal flow for a Glow model having K flow-steps can be illustrated as

$$\mathbf{z} = \mathbf{h}_0 \xrightarrow{\mathbf{g}'_{s,k}^{[1]}} \mathbf{h}_1 \xrightarrow{\mathbf{g}'_{s,k}^{[2]}} \mathbf{h}_2 \xrightarrow{\mathbf{g}'_{s,k}^{[3]}} \mathbf{h}_3 \dots \xrightarrow{\mathbf{g}'_{s,k}^{[K]}} \mathbf{h}_K = \mathbf{x}. \quad (19)$$

Each i^{th} flow-step is composed of the three layers as

$$\mathbf{f}'_{s,k}^{[i]} = \mathbf{f}'_{1s,k}^{[i]} \circ \mathbf{f}'_{2s,k}^{[i]} \circ \mathbf{f}'_{3s,k}^{[i]}. \quad (20)$$

In (20), $\mathbf{f}'_{1s,k}^{[i]}$ denotes the activation normalization layer, $\mathbf{f}'_{2s,k}^{[i]}$ denotes the invertible convolution layer, and $\mathbf{f}'_{3s,k}^{[i]}$ denotes the affine coupling layer. We refer to the *forward mapping* as the

transformation from the data space to the latent space, and the *reverse mapping* as the transformation from the latent space to the data space. The log-determinant is usually computed for the forward mapping at each of the constituent layers shown in (20). Each of the individual layers of a single flow-step are described as follows:

a) *Activation normalization*: It is assumed that an input vector is represented as \mathbf{x} and is of the shape $(N_b \times N_c \times N_s)$, where N_b denotes the batch-size, N_c denotes the number of channels and N_s denotes the samples. In the following equation, \mathbf{x} can be the starting input vector or the activation at any intermediate hidden layer in the flow network. The forward, reverse functions and log-determinant are

$$\begin{aligned} \mathbf{z} &= (\mathbf{x} - \boldsymbol{\mu}_{bias}) \odot \exp(\boldsymbol{\sigma}_{scale}^*) && \text{(Forward direction)}, \\ \mathbf{x} &= \mathbf{z} \odot \exp(-\boldsymbol{\sigma}_{scale}^*) + \boldsymbol{\mu}_{bias} && \text{(Reverse direction)}, \\ \mathbf{J}_1 &= \sum_{\text{channel dim.}} \boldsymbol{\sigma}_{scale}^* && \text{(Log-determinant)}. \end{aligned} \quad (21)$$

In (21), the values of $\boldsymbol{\mu}_{bias}$ and $\boldsymbol{\sigma}_{scale}$ are calculated as the mean and the logarithm of the standard deviation along each of the channel dimensions.

b) *Invertible 1×1 convolution*: The invertible 1×1 convolution is said to be a generalization of the permutation along the channel dimensions [12]. Any convolution operation can be thought of as a matrix multiplication. The forward, reverse and log-determinant calculation are shown in (22) as

$$\begin{aligned} \mathbf{z} &= \mathbf{W}\mathbf{x} && \text{(Forward direction)}, \\ \mathbf{x} &= \mathbf{W}^{-1}\mathbf{z} && \text{(Reverse direction)}, \\ \mathbf{J}_2 &= \log(|\det(\mathbf{W})|) && \text{(Log-determinant)}. \end{aligned} \quad (22)$$

c) *Affine coupling layer*: The affine coupling layer used in Glow is very similar to that in RealNVP. It consists of a neural network function $\mathbf{N}(\cdot)$ that is usually a *shallow* feed-forward network using convolutional layers. The computations and log-determinant calculation in the forward direction are

$$\begin{aligned} \mathbf{x}_a, \mathbf{x}_b &= \text{split}(\mathbf{x}), \\ \log(\boldsymbol{\sigma}), \boldsymbol{\mu} &= \mathbf{N}(\mathbf{x}_b), \\ \boldsymbol{\sigma} &= \exp(\log(\boldsymbol{\sigma})), \\ \mathbf{z}_a &= (\mathbf{x}_a + \boldsymbol{\mu}) \odot \boldsymbol{\sigma}, \\ \mathbf{z}_b &= \mathbf{x}_b, \\ \mathbf{z} &= \text{concat}(\mathbf{z}_a, \mathbf{z}_b), \\ \mathbf{J}_3 &= \sum_{\text{channel dim.}} \boldsymbol{\sigma}_{scale} && \text{(Log-determinant)}. \end{aligned} \quad (23)$$

The design as per (23) indicates that $\log(\boldsymbol{\sigma})$ is modelled as an logarithm of the inverse of the scaling parameter and $\boldsymbol{\mu}$ is modelled as negative of the translation parameter. The only difference in integrating the Glow model with the HMM comes in the formulation of the log-determinant similar to (16) for RealNVP flow. Using the individual determinant calculations

from (21), (22), (23), we can calculate $\log(|\det(\nabla \mathbf{f}_{s,k})|)$ as

$$\begin{aligned} \log\left(\left|\det\left(\nabla \mathbf{f}'_{s,k}\right)\right|\right) &= \sum_{l=1}^K \log\left(\left|\det\left(\nabla \mathbf{f}'_{s,k}^{[l]}\right)\right|\right) \\ &= \sum_{l=1}^K \left(\mathbf{J}_1^{[l]} + \mathbf{J}_2^{[l]} + \mathbf{J}_3^{[l]}\right). \end{aligned} \quad (24)$$

In (24), K refers to the number of *flow-steps* (without considering a multi-scale flow architecture i.e. assuming $L = 1$). So once the log-determinant is calculated, we can substitute the value and calculate the cost function as in (12).

E. Training algorithm

The essential steps of the learning procedure are described in pseudocode in Algorithm 1. The algorithm involves training a flow-based model using Expectation-Maximization (EM) algorithm [35] and Back-propagation. It involves two essential steps:

- The first step which involves calculation of the posterior distribution $p(\mathbf{s}, \mathbf{k} | \mathbf{x}; \mathbf{H}^{old})$, and the loss function $\mathcal{L}(\Phi; \mathbf{H}^{old})$ involving the parameters of the output distribution. However, unlike common generative models like GMM-HMM, the NMM-HMM models require to be trained before they can be used, without which the sample statistics would no longer be stationary. So, there is an inner loop to train the flow-models using mini-batch gradient descent with Adam [37]. Training continues until a convergence criterion is satisfied or the maximum number of iterations is reached. The convergence criterion \mathcal{C}^* was defined as

$$\mathcal{C}^* : \left[\left(\frac{|\Lambda_{current} - \Lambda_{prev}|}{|\Lambda_{prev}|} < \Delta \right) == \mathbf{True} \right]$$

If \mathcal{C}^* was satisfied for a chosen number of times (n^*) in succession, the training was stopped, else the model continued to be trained until the maximum number of training iterations (N_{max}) was reached. The convergence threshold (Δ) was decided carefully through experimentation.

- The second step which consists of updating the Markov-chain parameters \mathbf{q} , \mathbf{A} and the mixture model weights $\mathbf{\Pi}$ were updated using standard optimization techniques of HMM training.

These two steps are iteratively continued until the likelihood maximization reaches a local optimum.

III. EXPERIMENTS AND RESULTS

This section describes evaluation of NMM-HMM models (NVP-HMM and Glow-HMM) for phone recognition in comparison with GMM-HMM models.

Algorithm 1: Learning algorithm for training NMM-HMMs

Input: Dataset $\{\mathbf{x}^{(m)}\}_{m=1}^M$, initial model parameters
Result: Optimized model parameters: \mathbf{q} , \mathbf{A} , $\mathbf{\Pi}$, Φ
Set training parameters: learning rate (η), no. of samples in a mini-batch (R_b), max. number of training epochs (N_{max}), convergence threshold (Δ), no. of convergence steps (n^*), previous negative log-likelihood (Λ_{prev}), and current negative log-likelihood ($\Lambda_{current}$)
Initialization of models: \mathbf{H}^{old} , \mathbf{H} : where, $\mathbf{H} = \{\mathbf{q}, \mathbf{A}, S, p(\mathbf{x}|s; \Psi_s)\}$, and set $\mathbf{H}^{old} \leftarrow \mathbf{H}$
while \mathbf{H} has not converged **do**
 while ($n \leq N_{max}$) or \mathcal{C}^* is satisfied **do**
 Input a mini-batch from the dataset as $\{\mathbf{x}^r\}_{r=1}^{R_b}$, with batch-size R_b
 Compute the posterior $p(\mathbf{s}^r, \mathbf{k}^r | \mathbf{x}^r; \mathbf{H}^{old})$ and the cost function $\mathcal{L}(\Phi; \mathbf{H}^{old})$ as per (17) or (24) (depending on the type of flow model)
 Compute $\partial \Phi$ by optimising $\mathcal{L}(\Phi; \mathbf{H}^{old})$ (Mini-batch gradient descent using Adam [37])
 Update $\Phi \leftarrow \Phi + \eta \cdot \partial \Phi$
 Update $n \leftarrow n + 1$
 end
 Update $\mathbf{q} \leftarrow \arg \max_{\mathbf{q}} \mathcal{L}(\mathbf{q}; \mathbf{H}^{old})$
 Update $\mathbf{A} \leftarrow \arg \max_{\mathbf{A}} \mathcal{L}(\mathbf{A}; \mathbf{H}^{old})$
 Update $\mathbf{\Pi} \leftarrow \arg \max_{\mathbf{\Pi}} \mathcal{L}(\mathbf{\Pi}; \mathbf{H}^{old})$
 Update $\mathbf{H}^{old} \leftarrow \mathbf{H}$
end

A. Dataset, software and hardware

The experiments were carried out using the TIMIT dataset [15]. A phone is a distinct speech sound and is universal irrespective of the language under consideration. Phonemes characterize different phones in the context of a specific language. The TIMIT dataset is an English corpus, consists of spoken utterances labelled at the phoneme level. We use phones and phonemes interchangeably in this article. We use a ‘folded’ set of 39 phonemes as per standard convention [16]. The TIMIT dataset is comprised of two parts - a training set and a testing set, consisting of 4620 and 1680 utterances respectively. For creating noisy data, we used four varieties of additive noises from the NOISEX-92 dataset [38]. The code for the NMM-HMM models (NVP-HMM and Glow-HMM) was written using PyTorch and run using GPU support. The code for the GMM-HMM model was written in Python using a package called *hmmlearn* [39]. The codes are available at <https://github.com/anubhabghosh/genhmm>.

B. Feature extraction

From the speech utterances we compute MFCCs. We used 13-dimensional MFCCs computed frame-by-frame with 25ms window length and 10 ms frame shift. As per standard practice we also compute dynamic features delta (Δ) and double-delta

($\Delta\Delta$) coefficients from the MFCCs. So, in total, we use 39-dimensional feature vector concatenating MFCCs, Δ and $\Delta\Delta$. We performed standard mean and variance normalization.

C. Training of models

For the GMM-HMM and NMM-HMM models, the number of hidden states (s) was chosen between 3 and 5, depending on the mean length of speech signal (in terms of number of samples). The initial probability vector \mathbf{q} was initialised as a vector of zeros, except the first state having probability one. The state transition probability matrix \mathbf{A} was initialised as an upper triangular matrix. In case of GMM-HMMs, we use diagonal covariance matrices for the Gaussian components. For both NMM-HMM models, the conditional distribution for each state was formulated as in (2) and (3). We experimentally find the number of mixture components (N_{mix}) for both the models. We also explain some additional, model-specific terms for NMM-HMMs. For NVP-HMM, we refer to a pair of consecutive coupling layers described in (13) and (14) as a *flow block*. It was necessary to ensure that the signal in the l^{th} layer would be alternated in the $(l+1)^{th}$ layer, so that there is a mixing of the signals between consecutive coupling layers. For Glow-HMM, we define a similar concept in terms of the parameters K and L , which refers to the number of *flow-steps* and number of *layers of multi-scale flow* respectively. One *flow-step* in the Glow model consists of three layers: activation normalization, invertible convolution, and an affine coupling layer (similar to RealNVP flow) shown in (20). We avoided the use of multi-scale flows and chose the value of K through experiments.

We use the Adam optimizer for realizing mini-batch gradient descent in NMM-HMM models [37]. The learning rate of gradient descent was experimentally set to be $\eta_{NVP} = 4 \times 10^{-3}$ for NVP-HMM and $\eta_{Glow} = 1 \times 10^{-4}$ for Glow-HMM. Additionally, the learning rate was adaptively decreased by a step-decay to aid training for more number of iterations. The EM algorithm involved maximization of the log-likelihood (or minimization of the negative log-likelihood), and monitoring relative change of the log-likelihood as a convergence criterion (explained in Section II-E).

D. Performance results

We now show performance results of competing models for different training and testing conditions.

1) *Clean training and testing*: The performances of the NVP-HMM and Glow-HMM models were compared with the baseline GMM-HMM model for training and testing on clean data. The results are shown in Table I. The GMM-HMM model was trained and tested on the clean data, for varying number of mixture components, i.e. $N_{mix} = \{1, 3, 5, 10, 15, 20\}$, and the model with the best performance was chosen. It was found that test accuracy for GMM-HMM increased with more number of mixture components. Based on the results, we decide to use $N_{mix} = 20$ for all our GMM-HMM models.

We next consider NVP-HMM and it is found that $N_{mix} = 3$ provides an improvement than $N_{mix} = 1$. Further improvement by increasing number of components was challenging for us due to high computational complexity. Therefore we decide to use NVP-HMM using $N_{mix} = 3$ components. Comparing results in Table I, we found that the NVP-HMM model with $N_{mix} = 3$ mixture components outperforms the baseline GMM-HMM model with $N_{mix} = 20$ components. The performance improvement is 4.8%. Here we mention that the NVP-HMM based ML-classification provides a similar classification accuracy for TIMIT phone recognition in comparison with discriminative training based DBN that had 77% accuracy [27].

Similar experiments were also carried out for the Glow-HMM model. It was first necessary to decide upon the number of flow-steps (K) required for the Glow model. As Glow-HMM is computationally complex, we used a smaller set of training data consisting of 5 classes (or phonemes) to decide the value of K . The result is shown in Table II. and we decide to use $K = 12$ flow-steps and $L = 1$ (no multi-scale flow). Also, it is worth noting that the accuracy values are high because it is relatively easier for the model when we use five classes instead of 39 classes. In order to achieve good performance, the use of *weight normalization* on the weights of the convolutional layer in the affine coupling layer of the Glow model in each flow-step was helpful, and also normalizing the activations by their euclidean norm. The weight normalization approach was inspired by its use in the *WaveGlow* model [40]. We believe that use of normalization schemes ensures better convergence towards desired negative log-likelihood values during training of Glow-HMM. Glow-HMM provided 76.3 % accuracy for 39 phones. This result is included in Table I for a comparison. Simulations for Glow-HMM were only carried out using $N_{mix} = 1$. For a higher number of mixture components, it was computationally expensive for us to perform further experiments.

While we have provided accuracy values in Table I, we also show other performance measures such as precision, recall and F1-score in Table III. We computed class-wise precision, recall and F1-scores and combined them into a weighted metric based on the number of samples in each class. Precision is computed as the ratio of the number of correct positive predictions in that class to the total number of positive predictions for the class. Recall (or sensitivity) measures the fraction of correct positive predictions to the number of actual positive cases, while F1-score is a weighted mean of precision and recall for the given class. The results show that the Glow-HMM has the largest precision, which reflects its ability to distinguish true and false positives. Also, the recall value of Glow-HMM is similar to NVP-HMM, which indicates that the NMM-HMM models are better at differentiating false negatives than GMM-HMM. But, the F1-score reflects that NVP-HMM is equally capable like Glow-HMM for clean data.

TABLE I: Test accuracy (in %) for GMM-HMM and NMM-HMM models at varying number of mixture components (N_{mix}) in clean training and clean testing scenario

Model-Type	No. of mixture components (N_{mix})					
	$N_{mix}=1$	$N_{mix}=3$	$N_{mix}=5$	$N_{mix}=10$	$N_{mix}=15$	$N_{mix}=20$
GMM-HMM	62.3	66.7	68.5	70.8	71.9	72.8
NVP-HMM	76.7	77.6	-	-	-	-
Glow-HMM	76.3	-	-	-	-	-

TABLE II: Training and Test accuracy (in %) for Glow-HMM ($N_{mix} = 1, L = 1$) at varying number of flow-steps (K) using the subset of clean data where we use 5 classes

Accuracy (in %)	No. of flow-steps (K)			
	$K=4$	$K=8$	$K=12$	$K=16$
Training	90.6	94.9	96.9	96.1
Testing	89.7	93.9	96.0	95.89

TABLE III: Overall weighted Precision, Recall, and F1 Score (in %) for three models in the clean training and clean testing scenario

Model-Type	Weighted Metrics		
	Precision	Recall	F1-Score
GMM-HMM	75.4	72.5	73.2
NVP-HMM	77.5	77.0	76.6
Glow-HMM	78.6	76.4	74.8

2) *Clean training and noisy testing - robustness test:* We now report results of experiments for checking the robustness of the NMM-HMM models relative to GMM-HMM in a mismatched training-testing scenario. We trained all the three models using clean (noise-less) data and tested them on noisy data. We used four types of noises: white, pink, babble and high frequency channel (labelled as ‘hfchannel’) at different signal-to-noise-ratio (SNR) levels. We expect that a model robust to noise, would show a good performance as well as a graceful degradation in performance as the noise power increases (or equivalently SNR decreases). Table IV shows the results. We observe that the performances of NMM-HMM models are better than GMM-HMM. We also note that Glow-HMM shows a better robustness to noise than NVP-HMM as exhibited by the graceful degradation of performance drop.

3) *Noisy training and noisy testing - robustness test:* In this set of experiments, the GMM-HMM and NVP-HMM were trained using a dataset that is comprised of clean as well as noisy data, and then tested on varying noisy conditions. In the training dataset, the portion of the noisy data is generated by corrupting the clean data using white noise at 10dB SNR. Test results for varying SNR conditions (only white noise) are shown in Table V. The test performance was evaluated on the clean test set, as well as white noise corrupted test set at different SNR levels. It was hypothesized that there would be an increment in test accuracy for noisy conditions when compared to the respective results in Table IV. The reason is the models have now been trained on noisy data. The results of Table V are plotted in Fig. 2 and we observe an interesting trend. Performance of both the models increases with SNR at first, then drops and then further increases. There is a peak

closer to 10 dB SNR. The reason lies how the training and testing conditions match. Note that the training data has a portion that is corrupted by white noise at 10 dB SNR. From this experiment, it is clear that NVP-HMM is more robust than GMM-HMM.

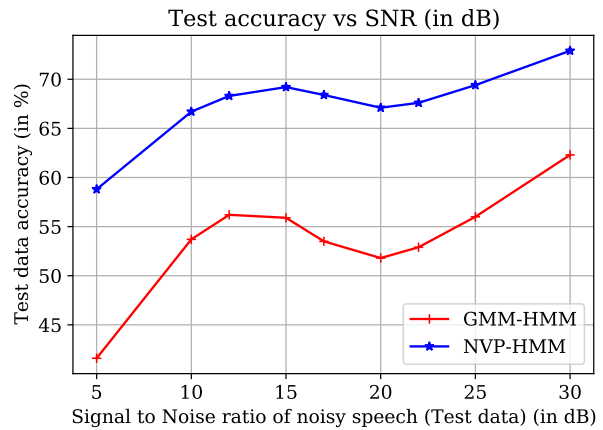


Fig. 2: Robust performance for noisy training. Both models were trained on a combination of clean data and noisy data where we used white noise at 10dB SNR. Then the test is performed using noisy test data where white noise is used at varying SNR values.

E. Improvements using decision fusion

While we achieved promising performance using generative models based ML classification in the previous experiments, we are still away from the state-of-the-art performance 86.2% accuracy [1]. Our main objective in this subsection is describing the engineering approach that can provide a performance close to state-of-the-art. Our hypothesis was that the different models behaved differently across phones (classes), and that developing an approach to combine their predictions would yield a better classifier. In this section, we describe our approach towards a *decision fusion (or voting)* scheme using GMM-HMM, NVP-HMM and Glow-HMM. We use traditional speech processing knowledge for grouping phonemes as per different categories such as Vowels, Semi-vowels, Plosives, etc. We hypothesized that this would help us to investigate whether some trends in performance can be observed based on the given category of phonemes and the statistical aspect of data availability for the particular class. Before explaining the scheme, we explain two aspects.

TABLE IV: Test accuracy (in %) for clean and noisy testing conditions, where we use clean training. The performance drop is shown in parenthesis with respect to the clean train and clean test scenario as in Table I

Clean data performance as a reference: GMM-HMM: 72.8, NVP-HMM: 77.6 and Glow-HMM: 76.3				
Type of Model	SNR levels for different kinds of noises			
	SNR = 25dB	SNR = 20dB	SNR = 15dB	SNR = 10dB
white noise				
GMM-HMM	55.6 (17.2)	46.8 (26.0)	36.8 (36.0)	27.9 (44.9)
NVP-HMM	67.1 (10.5)	60.0 (17.6)	49.4 (28.2)	37.7 (39.9)
Glow-HMM	67.9 (8.4)	61.4 (14.9)	53.6 (22.7)	45.8 (30.5)
pink noise				
GMM-HMM	59.9 (12.9)	51.9 (20.9)	42.3 (30.5)	32.2 (40.6)
NVP-HMM	69.3 (8.3)	61.7 (15.9)	48.6 (29.0)	33.7 (43.9)
Glow-HMM	69.4 (6.9)	62.9 (13.4)	54.9 (21.4)	45.9 (30.4)
babble noise				
GMM-HMM	65.7 (7.1)	59.3 (13.5)	49.3 (23.5)	37.4 (35.4)
NVP-HMM	70.7 (6.9)	65.8 (11.8)	56.2 (21.4)	42.3 (35.3)
Glow-HMM	72.4 (3.9)	68.9 (7.4)	63.1 (13.2)	53.8 (22.5)
hfchannel noise				
GMM-HMM	62.3 (10.5)	54.4 (18.4)	44.1 (28.7)	33.3 (39.5)
NVP-HMM	67.9 (9.7)	63.4 (14.2)	55.8 (21.8)	44.9 (32.7)
Glow-HMM	71.8 (4.5)	67.9 (8.4)	61.1 (15.2)	51.9 (24.4)

TABLE V: Test accuracy (in %) for GMM-HMM and NVP-HMM using noisy training

Model-Type	Clean	Diff. SNR levels for adding white noise								
		30dB	25dB	22dB	20dB	17dB	15dB	12dB	10dB	5dB
GMM-HMM	72.0	62.3	56.0	52.9	51.8	53.5	55.9	56.2	53.7	41.6
NVP-HMM	76.8	72.9	69.4	67.6	67.1	68.4	69.2	68.3	66.7	58.8

- *Sample ratio*: For a phone class, sample ratio is computed as

$$\text{Sample ratio} = \frac{\# \text{ tr. samples of the class}}{\# \text{ tr. samples in the largest class}}$$

This ratio helps us to understand the role of data availability for classification performance. Here ‘largest class’ means the ‘silence’ class for which the number of training samples is the highest among all classes.

- *Voting*: We use a majority voting principle. Among three models, if at-least two models are in agreement then we choose the agreed classification decision. Otherwise, all the three methods are in disagreement and then we choose a classification decision as a random choice among them. Note that voting is amenable to accommodate a growing number of classes.

Using this voting scheme, we were able to achieve an accuracy 86.6% on the test set, which is close to the state-of-the-art performance of 86.2% in [1]. We now show class-wise accuracy results for each model in Table VI. From the table, we observe a significant variation in performances of the three models. We show few examples. For ‘aa’ vowel, while GMM-HMM provides 69.9% accuracy, the Glow-HMM provides 98.38% accuracy. Then, for several diphthongs, such as ‘ow’, the Glow-HMM shows poor performance. For ease of visualization, we show variation in performance for vowels and fricatives in Fig. 3. This significant variation in performances is related to statistical diversity and helps the voting scheme to achieve 86.6% accuracy. The 86.6% accuracy is a significant improvement than the three models.

Our final experiment investigates the robustness of the voting based decision fusion approach. In this case, we have models that are trained using clean condition data and tested in noisy conditions. The results are shown in Table VII. In the table we compare the voting against the best performing model among the three models, which can be seen from the Table IV. We observe that the voting approach gracefully degrades with increasing noise power.

IV. CONCLUSIONS

In pursuit of explainability under maximum-likelihood classification and optimization, we show that it is possible to combine modern neural networks (normalizing flows), modern computational tools (Kaldi, Python and PyTorch), time-tested generative models for sequential data (HMMs), their machine learning based optimization (EM and mini-batch gradient search), and knowledge in domain specific signal processing (MFCCs in speech signal processing). The output of our endeavor can provide state-of-the-art performance.

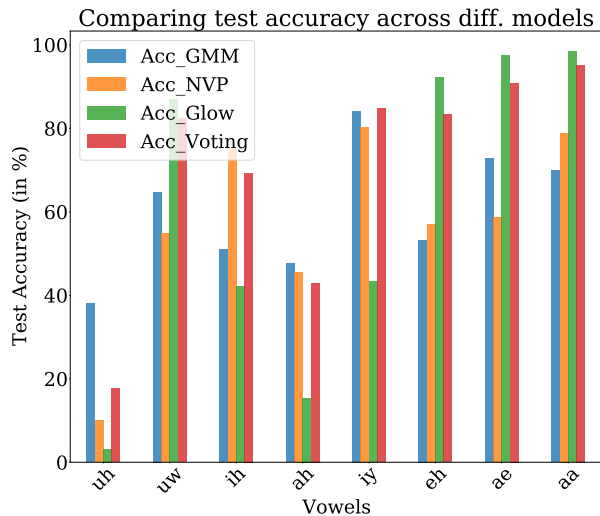
We deliberately refrain from using a discriminative approach in this article on the ground of explainability. For the same reason, we also avoid carefully crafted discriminative features. Still we are able to achieve 86.6% phone recognition accuracy whereas the latest result is 86.2% [1]. We refrain to claim that we are better by 0.4%. Our experience is that reporting of experimental results in machine learning with real data has some undue opaqueness and published results are typically in favour of authors, even though a reproducible research culture has been steadily growing.

TABLE VI: Test accuracy (in %) for phones using voting based decision fusion against GMM-HMM, NVP-HMM and Glow-HMM. We used clean training and clean testing scenario.

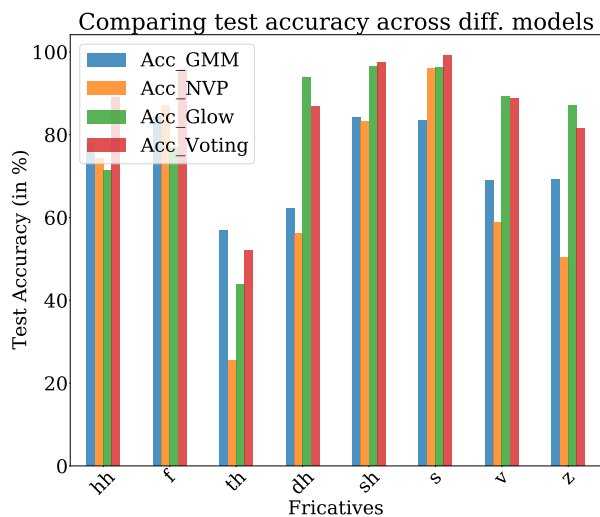
Index	Phoneme	Type	Sample ratio (%)	Acc. GMM - HMM	Acc. NVP - HMM	Acc. Glow - HMM	Acc. Voting
1	uh	V	1.49	38.01	9.95	3.17	17.65
2	uw	V	6.87	64.80	54.80	86.93	82.40
3	eh	V	10.75	53.19	57.08	92.15	83.33
4	ae	V	11.15	72.78	58.78	97.44	90.83
5	aa	V	16.75	69.90	78.90	98.38	95.11
6	ah	V	17.55	47.59	45.54	15.41	42.89
7	iy	V	19.39	84.13	80.30	43.43	84.69
8	ih	V	38.20	50.95	74.99	42.14	69.32
9	y	SV	4.78	76.18	64.67	49.68	77.44
10	w	SV	8.75	81.84	83.05	68.28	90.48
11	er	SV	15.21	76.49	80.11	84.33	92.62
12	r	SV	18.23	68.09	65.83	99.29	90.61
13	l	SV	18.83	65.21	84.99	99.04	96.33
14	ch	P	2.29	72.97	54.44	1.54	51.35
15	jh	P	3.37	69.09	63.17	75.54	82.80
16	g	P	5.57	71.18	68.23	33.65	70.24
17	b	P	5.59	74.28	66.04	64.92	83.15
18	p	P	7.15	67.90	77.22	76.48	88.35
19	d	P	9.51	68.21	71.61	91.23	90.73
20	t	P	12.11	62.82	72.66	86.95	87.41
21	k	P	13.54	77.28	83.86	62.88	88.89
22	ng	N	3.82	78.76	29.59	47.73	58.23
23	m	N	11.22	77.17	78.38	62.44	84.76
24	n	N	24.42	63.34	83.31	73.38	87.94
25	th	F	2.10	56.93	25.47	43.82	52.06
26	v	F	5.56	69.01	58.73	89.15	88.73
27	hh	F	5.89	77.79	74.21	71.31	88.97
28	f	F	6.18	84.43	87.17	76.86	95.72
29	sh	F	6.66	84.14	83.10	96.44	97.36
30	dh	F	7.77	62.09	56.24	93.76	86.76
31	z	F	10.53	69.29	50.35	86.96	81.62
32	s	F	20.85	83.52	96.10	96.29	99.20
33	oy	D	1.91	88.97	72.24	95.82	95.82
34	aw	D	2.03	70.83	25.00	2.78	32.87
35	ow	D	5.95	66.15	51.74	8.37	50.84
36	ey	D	6.36	80.52	76.92	11.54	75.43
37	ay	D	6.65	82.39	81.10	53.40	85.68
38	dx	C	7.55	86.49	78.72	95.64	97.13
39	sil	C	100.00	84.89	94.83	99.02	99.38
Performance for clean data training and testing (full set of 39 classes): GMM: 72.5%, NVP: 76.9%, Glow: 76.3%, Voting: 86.6%							
V - Vowels, SV- Semivowels, N - Nasals, D - Diphthongs, C - Closures, P - Plosives							

TABLE VII: Test accuracy (in %) for clean and various noisy conditions. We compare the performance of the best performing model between GMM-HMM, NVP-HMM and Glow-HMM with respect to the performance of the voting based decision fusion. We use the notation GMM/NVP/Glow to represent the best performing model among the three models.

Performance for clean data as a reference: GMM: 72.5, NVP: 76.9, Glow: 76.3, Voting: 86.6									
Type of Noise	SNR levels for different kinds of noises								
	25dB		20dB		15dB		10dB		Voting
	GMM / NVP / Glow	Voting	GMM / NVP / Glow	Voting	GMM / NVP / Glow	Voting	GMM / NVP / Glow		
white	67.9	76.5	61.4	66.5	53.6	54.1	45.8	42.6	
pink	69.4	80.1	62.9	71.9	54.9	59.9	45.9	45.4	
babble	72.4	83.2	68.9	79.7	63.1	71.7	53.8	57.4	
hfchannel	71.8	80.9	67.9	75.3	61.1	66.1	51.9	53.8	



(a) Vowels



(b) Fricatives

Fig. 3: Plot of test accuracy for different models - GMM-HMM (Blue), NVP-HMM (Yellow), Glow-HMM (Green) - against voting-based decision fusion (in Red) for vowels and fricatives

ACKNOWLEDGMENT

The authors would like to thank KTH Digital Futures Center for support amid a pandemic.

REFERENCES

- [1] M. Ravanelli, T. Parcollet, and Y. Bengio, "The Pytorch-kaldi Speech Recognition Toolkit," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2019-May, pp. 6465–6469, 2019.
- [2] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light Gated Recurrent Units for Speech Recognition," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 92–102, 2018.
- [3] M. Gales, S. Young *et al.*, "The application of hidden markov models in speech recognition," *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2008.

- [4] R. Plamondon and S. N. Srihari, "Online and off-line handwriting recognition: a comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63–84, 2000.
- [5] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *International Conference on Pervasive Computing*, Springer, 2004, pp. 1–17.
- [6] D. Anastassiou, "Genomic signal processing," *IEEE Signal Processing magazine*, vol. 18, no. 4, pp. 8–20, 2001.
- [7] M. I. Piecyk and A. C. McKinnon, "Forecasting the carbon footprint of road freight transport in 2020," *International Journal of Production Economics*, vol. 128, no. 1, pp. 31–42, 2010.
- [8] I. Kobyzev, S. Prince, and M. A. Brubaker, "Normalizing flows: Introduction and ideas," *arXiv preprint arXiv:1908.09257*, 2019.
- [9] D. Liu, A. Honoré, S. Chatterjee, and L. K. Rasmussen, "Neural network based explicit mixture models and expectation-maximization based learning," in *International Joint Conference on Neural Networks (IJCNN)*, Budapest, 2020.
- [10] A. Ghosh, A. Honoré, D. Liu, G. Eje Henter, and S. Chatterjee, "Robust classification using hidden markov models and mixtures of normalizing flows," in *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2020.
- [11] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP," *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2019.
- [12] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1×1 convolutions," *Advances in Neural Information Processing Systems*, vol. 2018-December, no. 2, pp. 10 215–10 224, 2018.
- [13] G. Papamakarios, T. Pavlakou, and I. Murray, "Masked autoregressive flow for density estimation," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, no. Nips, pp. 2339–2348, 2017.
- [14] D. Liu, A. Honoré, S. Chatterjee, and L. K. Rasmussen, "Powering hidden markov model by neural network based generative models," in *European Conference on Artificial Intelligence (ECAI)*, Spain, 2020.
- [15] C. Lopes and F. Perdigao, "Phone recognition on the timit database," *Speech Technologies/Book*, vol. 1, pp. 285–302, 2011.
- [16] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.
- [17] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.
- [18] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.
- [19] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *arXiv preprint arXiv:1912.10557*, 2019.
- [20] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, "A survey of the state of explainable ai for natural language processing," *arXiv preprint arXiv:2010.00711*, 2020.
- [21] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, *Explainable AI: interpreting, explaining and visualizing deep learning*. Springer Nature, 2019, vol. 11700.
- [22] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *arXiv preprint arXiv:1506.06579*, 2015.
- [23] S. A. Cadena, M. A. Weis, L. A. Gatys, M. Bethge, and A. S. Ecker, "Diverse feature visualizations reveal invariances in early layers of deep neural networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 217–232.
- [24] I. Lemhadri, F. Ruan, and R. Tibshirani, "Lassonet: Neural networks with feature sparsity," *arXiv preprint arXiv:1907.12207*, 2019.
- [25] G. Synneave, "wer_we_are," GitHub Repository, https://github.com/syhw/wer_are_we, 2015.
- [26] L. Tóth, "Phone recognition with hierarchical convolutional deep maxout networks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–13, 2015.
- [27] A.-r. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *Nips workshop on deep learning for speech recognition and related applications*, vol. 1, no. 9. Vancouver, Canada, 2009, p. 39.
- [28] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, no. 3, pp. 6645–6649, 2013.

- [29] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *Advances in Neural Information Processing Systems*, vol. 2015-January, pp. 577–585, 2015.
- [30] L. Lu, L. Kong, C. Dyer, N. A. Smith, and S. Renals, "Segmental recurrent neural networks for end-to-end speech recognition," *Interspeech 2016*, pp. 385–389, 2016.
- [31] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "Rasta-plp speech analysis technique," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1992, pp. 121–124.
- [32] S. Chatterjee and W. B. Kleijn, "Auditory model-based design and optimization of feature vectors for automatic speech recognition," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 6, pp. 1813–1825, 2010.
- [33] C. Koniaris, S. Chatterjee, and W. B. Kleijn, "Selecting static and dynamic features using an advanced auditory model for speech recognition," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4342–4345.
- [34] B. Logan, "Mel frequency cepstral coefficients for music modelling," in *Proc. International Symposium on Music Information Retrieval (ISMIR 2000)*, 2000.
- [35] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.
- [36] L. Dinh, D. Krueger, and Y. Bengio, "NICE: Non-linear independent components estimation," *3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings*, vol. 1, no. 2, pp. 1–13, 2015.
- [37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [38] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.
- [39] S. Lebedev, "hmmllearn," GitHub Repository, <https://github.com/hmmllearn/hmmllearn>, 2014.
- [40] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.