Information Theoretic Analysis of DNN-HMM Acoustic Modeling

Pranay Dighe, Student Member, IEEE, Afsaneh Asaei, Senior Member, IEEE, and Hervé Bourlard, Fellow, IEEE

Abstract—We propose an information theoretic framework for quantitative assessment of acoustic modeling for hidden Markov model (HMM) based automatic speech recognition (ASR). Acoustic modeling yields the probabilities of HMM sub-word states for a short temporal window of speech acoustic features. We cast ASR as a communication channel where the input sub-word probabilities convey the information about the output HMM state sequence. The quality of the acoustic model is thus quantified in terms of the information transmitted through this channel.

The process of inferring the most likely HMM state sequence from the sub-word probabilities is known as decoding. HMM based decoding assumes that an acoustic model yields accurate state-level probabilities and the data distribution given the underlying hidden state is independent of any other state in the sequence. We quantify 1) the acoustic model accuracy and 2) its robustness to mismatch between data and the HMM conditional independence assumption in terms of some mutual information quantities. In this context, exploiting deep neural network (DNN) posterior probabilities leads to a simple and straightforward framework to assess shortcomings of the acoustic model for HMM based decoding. This analysis enables us to evaluate the Gaussian mixture acoustic model (GMM) and the importance of many hidden layers in DNNs without any need of explicit speech recognition. In addition, it sheds light on the contribution of low-dimensional models to enhance acoustic modeling for better compliance with the HMM based decoding requirements.

Index Terms—automatic speech recognition, acoustic modeling, information theory, deep neural networks, conditional mutual information, low-rank and sparsity.

I. INTRODUCTION

Over the last 40 years, hidden Markov models (HMMs, and more recently DNN-HMM) have served as the backbone of virtually all large scale Automatic Speech Recognition (ASR) systems [1], [2], [3]. However, HMMs are built upon several major assumptions which are well known and understood, yet often shattered in the speech community [4], [5], [6].

More specifically, the HMM theory relies on the following assumptions. Firstly, the probability distribution associated to a hidden state depends only on that state. Therefore, the acoustic observation is conditionally independent of all the rest given the underlying hidden state. This assumption limits the scope of temporal dependency captured by HMMs. Secondly, HMM is a first-order Markov model, i.e., the probability of the Markov chain to be in a particular state at a time step depends solely on the previously visited state and nothing else. Furthermore, traditional HMM based sequence modeling requires a-priori definition of the state probabilistic distribution, which was often considered as a mixture of multivariate Gaussians in GMM based HMM systems.

In modern hybrid DNN-HMM architectures, the DNNs make no such assumption about the statistical distribution of the observations and directly estimate the state-specific posterior probabilities, possibly conditioned on some limited temporal context. Replacing GMM acoustic models by DNNs to achieve these hybrid speech recognition systems [4] has been the single largest source of ASR performance improvement in the last few years [7]. This leap in performance achieved by using DNN acoustic models necessitates the study of the following fundamental questions:

- Does DNN based acoustic modeling specifically fulfill the HMM assumptions better than GMMs?
- And if so, can we formally identify some properties desired in an acoustic model for improving ASR performance?

The present work is an attempt towards answering the above questions by deriving an information theoretic analysis framework to probe multiple facets of DNN-HMM speech recognition. This work adds to the growing body of research (summarised in Section II-B) which has analysed the contribution of DNNs in improving HMM-based ASR. In a broader sense, it addresses the following open questions regarding the role of any acoustic model that can be considered for HMM based ASR:

1) How to quantify the quality of the acoustic model in terms of its ability to make accurate predictions?
2) How to evaluate the robustness of the acoustic model towards the violation of HMM assumptions?

This work takes a novel information theoretic standpoint to gather evidence that the improved ASR performance due to DNN-HMMs is a result of more accurate acoustic modeling as well as better compliance with the HMM assumptions for state decoding. The core analysis in this paper depends upon conceptualising the traditional HMM formulation in a different manner. Instead of treating a continuous multi-dimensional acoustic feature (e.g. a MFCC vector) as an observation, we treat the acoustic model’s state prediction as a discrete observable feature emitted by the hidden HMM state. This simple trick facilitates the computation of some important information theoretic terms which are otherwise highly non-trivial to compute using acoustic features. These information theoretic measurements can directly quantify the quality of the acoustic model for HMM based ASR instead of relying on full-fledged ASR related measurements like frame classification accuracy or word error rate (WER).
Thus, the result of the present study yields novel quantitative evaluation measures of an acoustic model without the need to perform explicit ASR. It also elucidates how DNN leads to ASR improvements and where should the research and design strategy of acoustic models be focused on to target the limitations in HMM-based decoding.

The rest of the paper is organised as follows: Section II provides background on HMM based speech recognition and discusses the relevant prior research in detail. Section III introduces a speaking-listening perspective to the process of ASR using a novel z-HMM formulation. Section IV introduces our information theoretic analysis of the desired acoustic model properties using z-HMM formulation. Section V describes the experimental results and analysis of our findings and finally, Section VI draws the conclusions of our work and discusses the directions for future work. The important notations for the mathematical expressions are listed in Table I.

II. BACKGROUND AND PRIOR RESEARCH

Speech is a complex time-varying signal which is usually assumed as resulting from a piece-wise stationary stochastic process so that the observed signal can be modelled by a Hidden Markov Model (HMM). HMM is a probabilistic finite-state model in which the state-specific emission probability distribution is assumed to be independent of previous states and previous observations, and the transition probabilities follow a first-order Markovian structure.

A. Foundation of HMM based Speech Recognition

In a typical HMM based ASR framework, the hypothesised word sequence $W$ is estimated from the sequence of acoustic features $X = \{X_1, \ldots, X_T\}$, where $X_t$ is a standard acoustic feature (e.g., MFCC with first and second order time derivatives) at time $t$, as

$$\hat{W} = \arg \max_W \max_p P(W|X)$$

where $P(W)$ is the probability of word sequence $W$ estimated from language model and $p(X|W)$ is the likelihood of the acoustic sequence conditioned on the word sequence, estimated from the acoustic model (consisting here of a DNN, followed by a HMM). Marginalising over all possible (hidden) state sequences, $p(X|W)$ is then computed as:

$$p(X|W) = \sum_Q p(X|Q,W)P(Q|W)$$

$$\approx \max_Q \prod_{t=2}^{T} a_{Q_{t-1}Q_t} \prod_{t=1}^{T} p(X_t|Q_t)$$

TABLE I: Notations and their associated meaning.

| Notation | Indication |
|----------|------------|
| $Q = \{q_1, \ldots, q_K\}$ | Set of senones where $K$ is the total number of senones; each senone corresponds to a HMM state. |
| $Q_t \in Q$ | Hidden state random variable at time index $t$ which takes values from the set of senones $Q$. |
| $Q_1, \ldots, Q_T >$ | Sequence of HMM states underlying a utterance. |
| $X_t$ and $x_t$ | Acoustic feature random variable and its value $\in \mathbb{R}^d$ denoting the observation at time $t$. |
| $X = \{X_1, \ldots, X_T\}$ | Sequence of acoustic features observed for an utterance. |
| $Z_t \in \{P(Z_t = q|x_t) \ldots P(Z_t = q_k|x_t)\}$ | Random variable denoting the state predicted by the acoustic model at time $t$; it takes values from the set of senones $Q$. |
| $Z = \{Z_1, \ldots, Z_T\}$ | Sequence of states predicted by the acoustic model. |

where $Q = \{Q_1, \ldots, Q_T\}$ is the most probable state sequence obtained from Viterbi algorithm for decoding. $\pi_{Q_t}$ is the initial state probability, $a_{Q_{t-1}Q_t}$ is the state transition probabilities, and $p(X_t|Q_t)$ is the likelihood of speech frame $X_t$ given the hidden state $Q_t$. In Bayesian acoustic modelling (GMM-HMM), the frame likelihood $p(X_t|Q_t)$ is modelled by a mixture of Gaussian distributions whereas in a hybrid deep neural network system (DNN-HMM), data (scaled) likelihood is approximated using state posteriors (obtained at the output of the DNN) and HMM-state prior probabilities as

$$p(X_t|Q_t) \approx \frac{p(X_t|Q_t)}{p(X_t)} \frac{P(Q_t|X_t)}{P(Q_t)}$$

where the state posterior probabilities $P(Q_t|X_t)$ are computed by a DNN, while prior probabilities $P(Q_t)$ are estimated from the state frequency counts over the training data.

The best acoustic model for HMM decoding should yield accurate state-specific probabilities (ideally independent of previous frames and previous states) leading to correct classification of the hidden state $Q_t$ conditioned (only) on acoustic feature $X_t$. Moreover, the HMM structure requires that for a given hidden state $Q_t$, the acoustic observation $X_t$ is conditionally independent of all the past and future hidden states $Q_{-t}$ and observations $X_{-t}$, where $\sim t$ denotes time index other than $t$.

B. Prior Research

Building upon work initiated in the early ’90s [4], and fully exploiting the availability of larger amount of training data and processing power, DNNs are now recognised to outperform GMMs in HMM based ASR [7]. Several studies have been conducted to better understand the reasons behind superior performance of DNN acoustic models. We review the previous research findings to enlist some of these properties desired in an acoustic model for improving HMM based ASR.

1) Towards Better State Conditional Probabilities: A detailed discussion in [3], [8] argues that accurate sequence decoding using HMM requires the acoustic model to be structurally discriminative of the underlying classes and the model should capture their distinctive features. Furthermore, the acoustic model should be designed to preserve a maximum of information (maximum mutual information) between the input features and its associated HMM state.

Recent investigations on the key factors contributing to the success of DNNs highlighted their invariant representation learning power for class discrimination [7, 9]. Unlike the generative GMM models, DNNs derive discriminative representations of the data by applying non-linear transforms...
through multiple hidden layers. In [10], it was found that the individual neurons in each of the DNN hidden layers learn to be selectively active in different ways towards distinct phone patterns. At the same time, information irrelevant to phonetic discrimination such as gender are discarded by the deeper (closer to the output) layers. It was also confirmed analytically in [11] that DNNs are significantly better at phone classification compared to GMMs and, although robustness against unseen noise and data/channel mismatch is a challenge, they still outperform GMMs in these conditions. Finally, it has been shown theoretically and experimentally that a DNN alleviates the need for an explicit probability distribution to model the data, and its data-driven discriminative approach leads to more accurate modeling of the underlying class (HMM states) distribution.

In short, the first lesson learnt so far from prior research attributes DNNs success partly due to the accurate estimation of the state-specific probabilities and better discrimination of the boundaries, resulting in superior HMM state-level classification results.

2) Acoustic Model Meets the Markovian Structure: Another body of works investigating the sources of the data-model mismatch are dedicated to studying the effects of HMM structural hypothesis on ASR failures. In particular, the conditional independence assumption of HMM is often acknowledged as the number one limiting factor resulting in poor ASR performance and lack of robustness [12].

Natural speech exhibits strong temporal correlations and contextual dependencies. This correlation is partly present in the acoustic features. HMM conditional independence assumption requires that the acoustic features associated with a specific sub-word (senone) state are independent of the past and future states and acoustic features. To test this hypothesis, earlier works [13], [6] replaced real speech data with synthetic data which strictly follow HMM assumptions. When conditional independence assumptions are strictly followed by the synthetic data, the ASR performance was found to be nearly perfect even using GMM-HMM architecture.

Along this line, [12], [14] have studied how DNNs cope with violation of HMM assumptions. It was shown that using many hidden layers in DNNs yields acoustic models less sensitive to the contextual dependencies. Each hidden layer successively makes the system more robust towards the contextual dependencies existing in the real speech data, hence resulting in better ASR.

A different approach to tackle the limitation of conditional independence assumption relies on alternative architectures that can capture longer temporal dependencies. For instance, the segmental models for stochastic modelling of speech [15] are employed to model a long duration of a speech segment as a unit instead of the frame-wise modeling procedure. In segmental models, the HMM conditional independence assumptions are not enforced within a segment, thus reducing the data-model mismatch. More recently, end-to-end ASR systems like attention-based mechanisms using long short term memory implementation of the recurrent neural networks (LSTM-RNNs) [16], [17], [18] are notable efforts in this direction.

The present work aims to theoretically quantify the contributions of DNNs and their impact on ASR performance to address the limitations imposed by HMM assumptions. Hence, building on the conclusions of earlier analytic studies and research objectives, we will show here that an important reason for the success of DNNs is their lower sensitivity to the contextual dependencies existing in the data which results in better fulfilment of the HMM conditional independence requirement.

In this paper, we consider the two desired properties of the acoustic model (discussed above and in Section II-B1), namely, the state conditional probabilities must be (1) accurate and (2) independent of each other.

C. Our Approach and Contributions

This work presents a novel information theoretic framework to conduct quantitative analysis of acoustic models for HMM-based ASR. The motivation is to understand and quantify what makes an acoustic model superior to others. Specifically, the proposed method is based on analysing the state conditional posterior probabilities generated by an acoustic model to evaluate their accuracy and compliance with the HMM assumptions.

One limitation of the previous studies in identifying the key reasons for the success of DNN based acoustic modeling is that they mostly rely on empirical evidences such as phone classification errors and ASR accuracies for validation of their hypotheses. In contrast, a methodology to quantify the desired acoustic model properties without performing ASR can measure the deficiencies in disjoint aspects. The present work attempts to address this need through a simple and straightforward information theoretic analysis framework.

The proposed solution casts ASR as a communication channel where the input state probabilities convey the information about the output state sequence. ASR performance is expected to be better when this channel has a high information capacity. We quantify the amount of transmitted information and the mismatch between the acoustic model and the HMM assumptions using information theoretic notion of mutual information. We demonstrate that higher capacity (accurate state posterior probabilities) and lower conditional mutual information (fulfilment of HMM assumptions) leads to improvement in ASR performance.

Unlike previous mutual information estimation using GMM that requires expectation maximization (EM) to learn a joint probability distribution [19], the proposed method directly uses the DNN based acoustic models in a simple and efficient information calculation procedure. Along this line, we introduce a novel z-HMM framework in Section III which facilitates quantifying the mutual information and leads to a novel perspective to ASR formulation.

As a use case, we apply the proposed evaluation framework to measure the effect of sparse and low-rank projections in improving DNN based acoustic modeling. It was shown in [20] that DNN based posterior probabilities live in a union of low-dimensional subspaces that can be characterized using
sparse [21] and low-rank representations [22], [23], [24]. Using the information theoretic analysis provided in this work, we are able to identify the specific contribution brought in by sparse and low-rank enhancements to the DNN acoustic models.

The next section introduces a speaking-listening perspective to the process of ASR.

III. SPEAKING-LISTENING z-HMM PERSPECTIVE

The graphical model for a traditional speech recognition HMM (in Figure 1(a)) as discussed in Section II consists of a sequence of hidden states which emit observable acoustic features at each time step. In terms of random variables, the hidden state $Q_t$ underlies the generation of feature $X_t$. Acoustic feature $X_t$ can take a value $x_t \in \mathbb{R}^n$ and state $Q_t$ can take a value $q_t \in \mathbb{Q}$ from the set of senones.

For the sake of the proposed analysis of acoustic models in HMM based ASR, we introduce a novel $z$-HMM formulation here, as follows. The acoustic model estimates the posterior probabilities of senone classes based on the observed acoustic feature $X_t$. This probabilistic prediction about the value taken by the random variable $Q_t$ (denoting the hidden state) is distinguished from the hidden state itself and is used to define a new random variable $Z_t$ which is now considered as an observable feature. The graphical model in Figure 1(b) depicts this process. We call this model as $z$-HMM since the emission from the hidden state $Q_t$ is the discrete random variable $Z_t$. Probability distribution over $Z_t$ is conditioned over $X_t$ and is given by the acoustic model in form of the posterior feature $Z_t$.

In that framework, we thus introduce distinct interpretations for the HMM’s hidden state and the acoustic model’s prediction as separate random variables corresponding to

- $Q_t$: Speaking random variable
- $Z_t$: Listening random variable

where the speaker intends the production of senone $Q_t$ at time $t$. The speech signal acoustic feature $X_t$ serves as the information bearing medium through which the listener infers the senone $Z_t$ in a probabilistic manner using the posterior probabilities $P(Z_t = q_k | X_t = x_t)$ for all $q_k \in \mathbb{Q}$. The task of recognition is to find the most likely hidden state sequence $Q$ given the probabilities of $Z_t$’s over $t$. Thus, we consider ASR as a communication channel (Figure 1(c)) where the input is the sequence of listening random variable $Z$, obtained from the acoustic model (e.g. DNN), and the output is the sequence of speaker random variable $Q$. From this perspective, $z$-HMM can be interpreted as a joint speaking-listening HMM where the speaking process is represented by the underlying sequence $Q$ leading to the observation $X$ and the listening process inferred from $X$ is represented by $Z$.

To complete the definition of $z$-HMM, we define the conditional posterior probabilities of $z$-HMM hidden states as follows:

\[
\forall q_k \in \mathbb{Q}; P(Q_t = q_k | Z_t = q_k) := P(Z_t = q_k | X_t = x_t) \quad (4)
\]

\[
P(Q_t = q_k | X_t = x_t) \quad (5)
\]

where equality (4) is by definition of $z$-HMM and the equality (5) is due to the way we defined random variable $Z_t$. In other words, a one-to-one relationship holds between the hidden states and the senone observation [25]. The other posterior probabilities given by acoustic model are distributed according to the one-to-one mapping (disjoint probabilities) and they are considered irrelevant to the current state probability for a particular senone observation.

It is important to note here that we do not actually have access to exact value taken by any particular $Z_t$; what we have is only the acoustic model’s probabilistic prediction about the random variable $Z_t$ taking values from the set of senones $\mathbb{Q}$ conditioned on the intermediate feature $X_t$.

In the next section, we present the information theoretic analysis of the acoustic models based on $z$-HMM formulation described above.

IV. INFORMATION THEORETIC ANALYSIS

In the context of $z$-HMM described above, the desired properties of the acoustic model, namely, (1) high accuracy of the state conditional posterior probabilities and (2) compliance with the HMM assumptions are quantified as...
(i) $I(Z_t; Q_t)$: the mutual information between the observed feature $Z_t$ and the underlying hidden state $Q_t$, and
(ii) $I(Z_t; Q_{t-1}|Q_t)$: the mutual information between the feature $Z_t$ and the former state $Q_{t-1}$ if the current hidden state $Q_t$ is known, respectively. The notion of mutual information is defined in Section IV-A. We will explain the relation between the above quantities and the desired acoustic model properties in Section IV-B.

A. Mutual Information

In information theory, the mutual information of two random variables quantifies the information conveyed about one random variable, by the other random variable. This concept is defined through the notion of entropy which measures the quantity of information held in a random variable [26].

For a discrete random variable $A$ which takes values $a \in A$, the entropy $H(A)$ is defined as

$$H(A) = -\sum_{a \in A} P(A = a) \log(P(A = a))$$  \hspace{1cm} (6)

Accordingly, the conditional entropy $H(A|B)$ of a random variable $A$ given another random variable $B$, which takes values $b \in B$, is defined as

$$H(A|B) = \sum_{b \in B} P(B = b) H(A|B = b)$$

$$H(A|B = b) = -\sum_{a \in A} P(A = a|B = b) \log(P(A = a|B = b))$$  \hspace{1cm} (7)

Entropy quantifies the uncertainty of a random variable; thereby, it increases as the uncertainty about the underlying values grows or $p(A)$ and $p(B)$ approach a uniform distribution.

Mutual information $I(A; B)$ between two random variables $A$ and $B$ is the measure of mutual dependence between the two variables. It quantifies reduction in uncertainty of $A$ due to the knowledge of $B$, and vice versa. It is defined as

$$I(A; B) = H(A) - H(A|B) = H(B) - H(B|A)$$  \hspace{1cm} (8)

Accordingly, conditional mutual information is defined as

$$I(A; B|C) = H(A|C) - H(A|B, C)$$  \hspace{1cm} (9)

B. Desired Acoustic Model Properties

The $z$-HMM ASR communication channel discussed in Section III is most efficient when 1) the observed input feature $Z_t$ and the underlying hidden state $Q_t$ which is to be decoded have the highest mutual information $I(Z_t; Q_t)$, and 2) the HMM conditional independence assumption is satisfied so that the mutual information between the feature and the former hidden state is minimized if the current hidden state is known, i.e. $I(Z_t; Q_{t-1}|Q_t)$ approaches zero. Thus, we discuss the following two properties of an ideal acoustic model for the best classification and compliance with the HMM structure:

1) High Information Transmission Capacity ($P_1$): To maximize the amount of information transmitted by the acoustic model for state decoding, it is desired to have a high mutual information between the feature $Z_t$ and the underlying state $Q_t$ expressed as

$$I(Z_t; Q_t) = H(Z_t) - H(Z_t|Q_t) \quad \forall t \in \{1, \ldots, T\}$$  \hspace{1cm} (10)

In an ideal scenario, no uncertainty should be left in the acoustic model’s prediction by revealing the underlying hidden state (i.e. $H(Z_t|Q_t)$ should be zero) because of the one-to-one mapping between $Q$ and $Z$. But, due to variability in the intermediate acoustic features $X_t$ and correlations between senone states, the acoustic model’s prediction $Z_t$ is not deterministic and the probabilities $P(Z_t|Q_t)$’s are not binary, leading to a non-zero value of $H(Z_t|Q_t)$ (more in Section V-F). We rely on $I(Z_t; Q_t)$ as the measure of the information transmitted by the acoustic model in the ASR communication channel [27]. Due to the data-processing inequality [26],

$$I(Z_t; Q_t) \leq I(X_t; Q_t)$$

where equality is achieved when $Z_t$ is the sufficient statistic of features $X_t$. Maximizing $I(Z_t; Q_t)$ thus ensures deriving highly informative features $Z_t$ from $X_t$ to transfer maximum information about $Q_t$ for state decoding.

2) First-order Markovian HMM Structure ($P_2$): It is desired that the acoustic model yields robustness to the HMM conditional independence assumption that is often violated by the speech acoustic features. In other words, the feature $Z_t$ emitted by an underlying state $Q_t$ should be independent of the past and future observations and states if the current state $Q_t$ is given. This property can be expressed as

$$Z_t \perp \{Q_{-t}, Z_{-t}\} | Q_t$$  \hspace{1cm} (11)

In this work, we confine our analysis of conditional independence only to the preceding hidden state $Q_{t-1}$ although our algorithmic approach is generally applicable to any order of dependency computation. We quantify the following condition:

$$Z_t \perp Q_{t-1} | Q_t$$  \hspace{1cm} (12)

To measure the amount of mutual dependence between $Z_t$ and $Q_{t-1}$, we deploy conditional mutual information as

$$I(Z_t; Q_{t-1}|Q_t) = H(Z_t|Q_t) - H(Z_t|Q_{t-1}, Q_t)$$  \hspace{1cm} (13)

In an ideal scenario, $I(Z_t; Q_{t-1}|Q_t) = 0$ indicates that the acoustic model fulfills the first-order Markovian requirement for HMM based decoding. Hence, the issue of data-model mismatch due to the long temporal correlations would be perfectly alleviated by the ideal acoustic model (DNN).

C. Computational Procedure using Posterior Features

In this section, we develop the procedure to quantify the desired properties $P_1$ and $P_2$ by using the posterior probability features generated by the acoustic model. This approach relies on DNN’s capability as an acoustic model to directly generate posterior probabilities for underlying senone classes. Assuming that the acoustic models are trained on some transcribed training data, we perform this analysis on a separate development dataset (which has transcription available) because
we need the ground truth based forced senone alignments to compute the various information theoretic terms explained above. Test data is not used for this analysis.

To measure $P_1$ and $P_2$, the following quantities must be computed and used for calculation of the mutual information measures in (10) and (13):

\[
\{H(Z_t), \ H(Z_t|Q_t), \ H(Z_t|Q_{t-1})\} \quad (14)
\]

The above entropy terms in turn require computation of the following probabilities:

\[
\{P(Z_t), \ P(Z_t|Q_t), \ P(Z_t|Q_{t-1})\} \quad (15)
\]

The acoustic model yields posterior features $z_t$ as:

\[
z_t = [P(Z_t = q_1|X_t = x_t) \ldots P(Z_t = q_K|X_t = x_t)]^T = P(Z_t|X_t = x_t)
\]

as probability distribution for the random variable $Z_t$ conditioned on the acoustic feature $X_t$. We generate the posterior features $z_t$’s and forced alignments $Q$ for all the utterances in the development data.

To compute $P(Z_t)$, we consider all the feature frames in a Monte Carlo method and approximate $P(Z_t)$ as the average posterior probability by marginalization over $X_t$:

\[
P(Z_t) = [P(Z_t = q_1) \ldots P(Z_t = q_K)]^T 
\approx \frac{1}{N} \sum_{t=1}^N P(Z_t|X_t = x_t) P(X_t = x_t) = \frac{1}{N} \sum_{t=1}^N z_t
\]

where $N$ is the total number of frames in the data. A uniform probability distribution is assumed for the acoustic features $X_t$. Given a large sample size in analysis of the large vocabulary ASR, ensemble averaging yields a reliable estimate.

To obtain the state conditional probabilities $P(Z_t|Q_t = q_k)$, we consider only the frames aligned to the state $Q_t = q_k$ in the forced alignments for marginalization over acoustic features.

Thus, we have

\[
P(Z_t|Q_t = q_k) 
\approx \frac{1}{N_{q_k}} \sum_{t s.t. Q_t = q_k} z_t
\]

where $N_{q_k}$ is the number of frames aligned to senone $q_k$.

In a similar manner, we compute $P(Z_t|Q_{t-1} = q_k, Q_{t-1} = q_{k'})$ by considering only those frames that are aligned to state $Q_t = q_k$ and the preceding frame is aligned to state $Q_{t-1} = q_{k'}$:

\[
P(Z_t|Q_t = q_k, Q_{t-1} = q_{k'}) 
\approx \frac{1}{N_{q_k,q_{k'}}} \sum_{t s.t. Q_t = q_k} \sum_{Q_{t-1} = q_{k'}} z_t
\]

where $N_{q_k,q_{k'}}$ is the number of frames aligned to senone $q_k$ such that preceding frame is aligned to senone $q_{k'}$.

The steps to compute the entropies in (14) and calculation of the required mutual information quantities from probabilities computed in (17), (18) and (19) are listed in Algorithm 1. The state prior probabilities and state transition probabilities involved in computations shown in Algorithm 1 are obtained by the frequency count approach using the ground truth forced alignment. Quality of the acoustic model used for ASR is measured based on a high value of $I(Z_t; Q_t)$ ($P_1$) and a low value of $I(Z_t; Q_{t-1}|Q_t)$ ($P_2$).

V. NUMERICAL EVALUATION AND ANALYSIS

Experiments are conducted on the challenging AMI corpus [28], consisting of conversational speech (in smart meeting rooms) in accented English. We use different acoustic models
TABLE II: Information theoretic analysis of different acoustic models to evaluate properties $P_1$ and $P_2$: Higher mutual information $I(Z_t; Q_t)$ indicates higher accuracy of the acoustic model and lower conditional mutual information $I(Z_t; Q_{t-1}|Q_t)$ shows compliance with the HMM conditional independence assumption. Information theoretic analysis is done on AMI-dev set and AMI-test set is used for evaluating ASR performance; last column shows word error rate (WER, in %). $x$HL denotes number of hidden layers and EP denotes equal parameters to the baseline DNN with 4 hidden layers.

| Acoustic Model | $H(Z_t)$ | $H(Z_t|Q_t)$ | $H(Z_t|Q_{t-1},Q_t)$ | $I(Z_t; Q_t)$ | $I(Z_t; Q_{t-1}|Q_t)$ | AMI-test WER% |
|----------------|----------|---------------|----------------------|--------------|-----------------------|--------------|
| GMM            | 9.826    | 5.346         | 5.078                | 4.480        | 0.268                 | 42.9         |
| DNN(-4HL)      | 9.547    | 4.468         | 4.262                | 5.079        | 0.206                 | 32.4         |
| LogReg(-0HL)   | 9.743    | 5.594         | 5.418                | 4.149        | 0.176                 | 52.0         |
| DNN-1HL        | 9.615    | 4.804         | 4.603                | 4.811        | 0.202                 | 36.9         |
| DNN-2HL        | 9.584    | 4.632         | 4.430                | 4.952        | 0.202                 | 34.5         |
| DNN-3HL        | 9.568    | 4.535         | 4.330                | 5.033        | 0.205                 | 32.8         |
| DNN-1HL-EP     | 9.599    | 4.732         | 4.528                | 4.867        | 0.203                 | 36.1         |
| DNN-2HL-EP     | 9.570    | 4.584         | 4.380                | 4.986        | 0.204                 | 34.0         |
| DNN-3HL-EP     | 9.570    | 4.530         | 4.327                | 5.040        | 0.203                 | 33.0         |
| Forced Aligned | 8.840    | 0             | 0                    | 8.840        | 0                     | -            |

and evaluate properties $P_1$ and $P_2$. The acoustic models considered are based on standard Gaussian mixture model (GMM) or DNNs with different architectures. These are studied in Section V-B. Furthermore, we investigate the effect of sparse and low-rank enhancements, as proposed in [22], on DNN acoustic models in Section V-E.

A. Experimental Setup

AMI corpus contains recordings of spontaneous conversations in multiparty meeting scenario. We use recordings from individual head microphones (IHM) comprising of around 67 hours of train set, 9 hours of development (dev) set, and 7 hours test set. All acoustic models are trained using the same training data. 10% of the training data is used for cross-validation during DNN training. Kaldi toolkit [29] is used for training of GMM and DNN-HMM systems.

The input features are 39 dimensional Mel frequency Cepstral coefficients (MFCC) together with their first order ($\Delta$) and second order ($\Delta\Delta$) temporal derivatives, hence resulting in $39 \times 9 = 351$ dimensional input, and an output class space of dimension 4007, representing the senone probability vector space. A GMM-HMM system is used to get the ground truth based forced senone alignments over the train set and dev set. These alignments serve as the hidden state sequences $Q$ for our analysis.

Based on previous experiments, our best baseline DNN acoustic model has 4 hidden layers with 1200 nodes each and it is trained using hard targets from GMM-HMM alignment. DNN posteriors (probability distributions over $Z_t$’s) are obtained for the dev set using forward pass, whereas the GMM posteriors are computed as the scaled likelihoods using Gaussian mixture distributions learned from the training data.

The information theoretic analysis of different acoustic models is shown in Table II. The last row shows various quantities by treating forced senone alignments as binary posterior features. Hence, the entropy $H(Z_t)$ here refers to the entropy of the prior probabilities of senone classes and this row essentially depicts the most ideal values we could hope to achieve from an acoustic model.

B. Comparing DNN v/s GMM Acoustic Modeling

We compare the first two rows of Table II here. Our primary observation is that DNN based probabilities on $Z_t$ have lower entropy $H(Z_t)$ than GMM’s. This means that the level of uncertainty in DNN outputs is typically lower than the GMM posteriors.

Moreover, we observe higher mutual information $I(Z_t; Q_t)$ in DNN acoustic models which indicates more information transmission through the ASR communication channel (cf. Figure 1(c)) as compared to GMMs. Thereby, the capacity of the channel is higher and the DNN posteriors are more accurate in discrimination of the underlying senone classes. This confirms the well known better modeling capability of DNN as compared to GMM.

Furthermore, the HMM conditional independence criterion is better satisfied by the DNN acoustic model as the conditional mutual information between current observation and the former state if the current state is given, $I(Z_t; Q_{t-1}|Q_t)$, is lower for DNN as compared to GMM. When these models are used to perform ASR on test data, DNN performs significantly better than GMM as expected. Nonetheless, we can quantify the desired acoustic model properties individually regardless of the ASR results.

Since the information theoretic criteria are computed prior to decoding for ASR, this study essentially disentangles the contribution of the language model. The additional information conveyed by the language model can be quantified nevertheless by re-estimation of the frame level senone posteriors after a full fledged ASR decoding.

C. Effect of Increasing Depth in DNN Acoustic Models

A very interesting evaluation is to measure the contribution of increasing the number of DNN hidden layers. The results...
are listed in Tables I. We compare DNN architectures in a manner similar to the study in [12] where number of hidden layers are increased with or without keeping the total number of network parameters equal to the baseline DNN. DNN with 0 hidden layer is simply a logistic regression model with input layer connected directly to the output layer.

It is a striking observation that the deeper architectures always have higher mutual information \( I(Z_t; Q_t) \) leading to higher acoustic model accuracy. This trend is observed in both the cases- when number of parameters are equal and when they are not. On the other hand, we observe quite negligible effect of the depth of DNN on the robustness towards violation of HMM conditional independence assumption. Regardless of this, we find that all DNN architectures have lower value of \( I(Z_t; Q_{t-1}|Q_t) \) as compared to the baseline GMM acoustic model.

The zero hidden layer logistic regression network has the poorest performance in the ASR task despite having the lowest correlation between the features and the past state. This is partly explained by the very low mutual information \( I(Z_t; Q_t) \) between its predictions and the underlying senone states. We also note the high values of state conditional entropies \( H(Z_t|Q_t) \) and \( H(Z_t|Q_t, Q_{t-1}) \) for this model which essentially indicate highly inaccurate senone predictions rendering the model inferior to GMM and DNN acoustic models.

\[ \text{D. Sparse and Low-rank Acoustic Model Enhancement} \]

In [22], we modify the forward pass outputs of the baseline DNN using (1) principal component analysis (PCA) based low-rank reconstruction and (2) overcomplete dictionary based sparse reconstruction. We learn principal component matrix and dictionary for each senone separately. Then, low-rank or sparse reconstruction is done in a supervised fashion by picking the PCA matrix or dictionary corresponding to the ground truth senone label for reconstruction.

This idea is illustrated in Figure 2(a). PCA projects data onto a low-dimensional subspace whereas sparsity achieves the same goal by expressing data as a sparse linear combination of columns of an overcomplete dictionary. This process essentially computes the projection of DNN output posterior features on the correct senone subspace. In terms of information theory, the acoustic modeling component (Figure 2(b)) now consists of DNN acoustic model followed by an additional block of principal component transform or dictionary based sparse coding.

We observe in Table III that enhanced DNN models have lower entropies \( H(Z_t) \) and \( H(Z_t|Q_t) \) than baseline DNN. Also, the mutual information \( I(Z_t; Q_t) \), between hidden senone state \( Q_t \) and the observation \( Z_t \), increases significantly using low-rank and sparse enhancements. These observations support our claim that low-rank and sparse reconstruction enforce the posterior features to capture more information related to the underlying senone subspace. Specially, in the case of PCA based reconstruction, we observe that the mutual information \( I(Z_t; Q_t) \) is maximum.

The increase in mutual information \( I(Z_t; Q_t) \) from baseline DNN to low-rank and sparsity based models explicitly quantifies the additional information that is contributed by the enhancement process. Note that PCA and dictionary based reconstruction are able to exploit the global information about the senone subspaces and thus they ought to bring additional information to the DNN local estimates. This information is available in terms of global patterns within each senone’s posterior features, but it is not accessible to the baseline DNN during a local forward pass of individual acoustic frames. It is only through the supervised enhancement using principal components or an overcomplete dictionary, that we are able to augment this global information in local framewise posterior features.

Another interesting observation is that the conditional independence assumption is also better satisfied (low values of \( I(Z_t; Q_{t-1}|Q_t) \)) in case of low-rank and sparsity based reconstruction. We explain it as follows. The frames aligned with senone \( q_k \) in the forced alignment can appear in different contexts in the data (see Figure 3). They exhibit different contextual information in DNN posteriors due to different neighboring senones. This contextual information which is always present in the real data violates the conditional independence assumption of HMM and leads to compromise in...
TABLE III: Information theoretic analysis of different acoustic models to evaluate properties \( P_1 \) and \( P_2 \) using AMI dev set. ASR evaluation is done on AMI test set using soft target training based student DNN acoustic models.

| Acoustic Model               | \( H(Z_t) \) | \( H(Z_t|Q_t) \) | \( H(Z_t|Q_{t-1},Q_t) \) | \( I(Z_t;Q_t) \) | \( I(Z_t;Q_{t-1}|Q_t) \) | AMI-test WER% |
|------------------------------|--------------|------------------|--------------------------|----------------|--------------------------|---------------|
| DNN                          | 9.547        | 4.468            | 4.262                    | 5.079         | 0.206                    | 32.4          |
| Sparse-SoftTargets DNN       | 9.317        | 3.399            | 3.237                    | 5.918         | 0.163                    | 31.6 (Student)|
| LowRank-SoftTargets DNN      | 9.250        | 2.751            | 2.653                    | 6.499         | 0.098                    | 31.6 (Student)|

ASR performance. When posterior features are reconstructed using PCA or an overcomplete dictionary, all the frames of senone \( q_k \) are forced to lie on the common subspace which defines \( q_k \). By controlling the parameters of PCA and sparse reconstruction, we ensure that only the most important dynamics of the senone subspace are preserved during enhancement of posteriors. Context dependent information, which is local to an individual frame and does not appear in the global patterns of the subspace, is reduced after reconstruction. Thus, the enhanced posterior features fulfill the conditional independence assumption better than the posteriors before reconstruction.

A caveat here is that the sparse and low-rank enhancements are done in a supervised fashion on the dev set by using the knowledge of forced alignments. Since we do not have alignments available for the test data, we can not directly evaluate the expected superiority of our acoustic model enhancement approach by performing ASR on test data. Instead, we use a simple methodology based on teacher-student DNN framework (used in [22]) to evaluate the ASR performance of our approach on test data. This is investigated in the following section.

E. Learning Student Networks Using Sparse or Low-rank Soft Targets

Recently, in [22], [23], we proposed to train student DNN acoustic models using enhanced soft targets obtained from supervised sparse and low-rank projection. Enhanced soft targets are basically training data posteriors obtained from the forward pass of a DNN trained on binary labels and then reconstructed using PCA or sparse coding in a supervised manner as shown in Figure 2.

The main idea of the teacher-student DNN training is that a “student” DNN can be trained with soft targets (instead of the one-hot labels) obtained from a more sophisticated model serving as the “teacher”. In our approach, the student DNNs trained using enhanced soft targets try to learn 1) the function modeled by the baseline DNN as well as 2) the reconstruction transformation performed by PCA or sparse coding.

In Table III, we provide the ASR performance of the sparse and low-rank soft targets based student DNNs to compare with the baseline DNN. By training student DNN models, we circumvent the issue that sparse and low-rank reconstruction requires forced alignments on the data. The student DNNs are expected to learn the reconstruction transformations implicitly in their parameters and enhance the test data posteriors accordingly during the forward pass. As expected, both the student models outperform the baseline DNN.

Theoretical analysis provided in this work supports the importance of sparse and low-rank enhancements in improving DNN acoustic models which was confirmed in our previous works [21], [22], [23], [24] experimentally. We can now better predict the performance of an acoustic model in ASR communication channel based on its information capacity and ability to better fulfil the HMM assumptions.

F. Further Implications

The speaking-listening \( z \)-HMM perspective provides different conceptual insights towards ASR. Assuming that the senone subspaces are independent, i.e. there is no correlation or contextual dependencies, a state \( Q_t = q_k \) should be deterministically mapped to the binary observation \( Z_t = q_k \) with the probability of one. However, as shown in [21], [22], senone subspaces are correlated with each other and they have ranks higher than unity for real data.

These correlations can arise from common parent nodes in senone decision trees or be the result of contextual dependencies of senones with each other as shown in Figure 4(a) and 4(b). Due to these correlations, a speaking (hidden) state

![Fig. 3: Sparse and low-rank reconstruction enforces different posterior features of a senone class to lie on a common low-dimensional subspace. Reconstructed posterior features have reduced local contextual correlations and satisfy HMM’s conditional independence criteria better.](image-url)
Fig. 4: Correlation among senones (a) long input contexts used in DNNs and (b) due to acoustically similar root in decision trees. In z-HMM, hidden state $Q_t$ corresponding to senone $s_k$ should emit senones from set S which are correlated with $s_k$.

$Q_t = q_k$ does not necessarily emit the listening state $Z_t = q_k$ deterministically. Instead, it can emit senone classes other than $q_k$ with non-zero probabilities (Fig 4(c)) and the one-to-one mapping is a limiting assumption.

Building on the theories elaborated in Section [III] and discussion above, the KL-HMM approach in the prior work [30]. [25] can be seen as associating a distribution $P(Z_t|Q_t)$ to each state. A state-specific distribution enables capturing the senone correlations. It can also characterize a probabilistic listening process that can be easily adjusted/adapted for accented speech production or varied pronunciations.

The proposed information theoretic measures may also be applied in finding the optimal number of senones for higher capacity of the acoustic model. Complementarity of multiple features in a multi-stream architecture can be quantified and feature selection strategies can be devised accordingly. Moreover, we can evaluate the new paradigms for acoustic modeling relying on long short-term memory (LSTM) and recurrent neural network (RNN) deep learning architectures, and compare them with DNNs in terms of the desired acoustic model properties. In addition, the contribution of the language model in ASR decoding can be measured and evaluated independent of the acoustic model.

The bottom line is that this analysis can pinpoint the sources of failures and improvements in ASR and measure the distinct factors separately and independently. Beyond diagnosis of the ASR building blocks, analysis of failures in adversarial conditions due to the effect of recording channels and noise and interferences can be conducted to quantify and evaluate the amount of information loss for ASR decoding.

**VI. Conclusions**

We presented an information theoretic approach to compare the quality of different acoustic models in HMM based ASR. We quantified the information transferred through an ASR communication channel as it decodes sequence of hidden HMM states. Higher amount of information transferred through this channel indicates better modeling capability of the acoustic model and higher accuracy. The conditional independence assumption of HMM is also evaluated in terms of the conditional mutual information between the current observation and the previous state if the current state is given. Lower value of the conditional mutual information shows better compliance with the HMM structure. Our experimental analysis yields quantitative measurement for these different dimensions of the superiority of DNN based acoustic modelling over GMMs.

In addition, we measured the contribution of incremental increase in the depth of DNN in improving the quality of the acoustic modeling. Furthermore, we showed that low-rank and sparse reconstruction done by PCA transforms and over-complete dictionaries for sparse coding respectively can further improve DNN acoustic modeling by bringing in additional global information about the senone subspaces to the DNN local estimations.

This analysis provides a novel approach to evaluate the quality of an acoustic model prior to using it for ASR on unseen test data. It also leads to a novel perspective to HMM based speech recognition exploiting the DNN acoustic model predictions as probabilistic features for state decoding. Future work will focus on the comprehensive ASR formulation and information theoretic evaluation of the acoustic modeling along with the language modeling component.

**Acknowledgements**

The research leading to these results has received funding from by Swiss NSF project on “Parsimonious Hierarchical Automatic Speech Recognition and Query Detection (PHASERQUAD)” grant number 200020-169398.

**References**

[1] Frederick Jelinek, “Continuous speech recognition by statistical methods,” Proceedings of the IEEE, vol. 64, no. 4, pp. 532–556, 1976.

[2] Lawrence R Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, 1989.

[3] Frederick Jelinek, Statistical methods for speech recognition, MIT press, 1997.

[4] Herve A Bourlard and Nelson Morgan, Connectionist speech recognition: a hybrid approach, Springer Science & Business Media, 1994.

[5] Jeff A Bilmes, “What HMMs can do,” IEICE Transactions on Information and Systems, vol. 89, no. 3, pp. 869–891, 2006.

[6] Dan Gillick, Larry Gillick, and Steven Wegmann, “Don’t multiply lightly: Quantifying problems with the acoustic model assumptions in speech recognition,” in IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2011, pp. 71–76.
[7] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, et al., “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, 2012.

[8] Jeff A. Bilmes, “What HMMs can’t do,” in ATR Workshop, Invited paper and lecture, 2004.

[9] Yoshua Bengio, “Learning deep architectures for ai,” Foundations and trends® in Machine Learning, vol. 2, no. 1, pp. 1–127, 2009.

[10] Tasha Nagamine, Michael L. Seltzer, and Nima Mesgraran, “Exploring how deep neural networks form phonemic categories.,” in INTERSPEECH, 2015, pp. 1912–1916.

[11] Dong Yu, Chaojun Liu, Yifan Gong, Yan Huang, “A Comparative Analytic Study on the Gaussian Mixture and Context Dependent Deep Neural Network Hidden Markov Models,” in Interspeech 2014, 2014.

[12] Suman Ravuri and Steven Wegmann, “How neural network depth compensates for hmm conditional independence assumptions in dnn-hmm acoustic models,” Interspeech 2016, pp. 2736–2740, 2016.

[13] Don McNallaster, Larry Gillick, Francesco Scattone, and Mike Newman, “Studies with fabricated switchboard data: Exploring sources of model-data mismatch,” in Proceedings of DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[14] Dan Gillick, Steven Wegmann, and Larry Gillick, “Discriminative training for speech recognition is compensating for statistical dependence in the hmm framework,” in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2012, pp. 4745–4748.

[15] M. Ostendorf, V. V. Digalakis, and O. A. Kimball, “From hmm’s to segment models: a unified view of stochastic modeling for speech recognition,” IEEE Transactions on Speech and Audio Processing, vol. 4, no. 5, pp. 360–378, Sep 1996.

[16] Hasim Sak, Andrew W. Senior, and Françoise Beaufays, “Long short-term memory recurrent neural network architectures for large scale acoustic modeling.,” 2014.

[17] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Yoshua Bengio, et al., “End-to-end attention-based large vocabulary speech recognition,” pp. 4945–4949, 2016.

[18] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 2016, pp. 4960–4964.

[19] Jeff A. Bilmes, “Maximum mutual information based reduction strategies for cross-correlation based joint distributional modeling,” in Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on. IEEE, 1998, vol. 1, pp. 469–472.

[20] Pranay Dighe, Afsaneh Asaei, and Hervé Bourlard, “Sparse modeling of neural network posterior probabilities for exemplar-based speech recognition,” Speech Communication, vol. 76, pp. 230–244, 2016.

[21] P. Dighe, G. Luyet, A. Asaei, and H. Bourlard, “Exploiting low-dimensional structures to enhance dnn based acoustic modeling in speech recognition,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 2016, pp. 5690–5694.

[22] Pranay Dighe, Afsaneh Asaei, and Hervé Bourlard, “Low-rank and sparse soft targets to learn better dnn acoustic models,” 2017.

[23] Pranay Dighe, Afsaneh Asaei, and Hervé Bourlard, “Exploiting eigen-posteriors for semi-supervised training of dnn acoustic models with sequence discrimination,” in Proceedings of Interspeech, 2017.

[24] Gil Luyet, Pranay Dighe, Afsaneh Asaei, and Hervé Bourlard. “Low-rank representation of nearest neighbor phone posterior probabilities to enhance dnn acoustic modeling,” in Interspeech, 2016.

[25] Marzieh Razavi, Ramya Rasipuram, and Mathew Magimai-Doss, “On modeling context-dependent clustered states: Comparing hmm/gmm, hybrid hmm/ann and kl-hmm approaches,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2014, pp. 7659–7663.

[26] Thomas M Cover and Joy A Thomas, “Entropy, relative entropy and mutual information.” 1991.

[27] Ohad Shamir, Sivan Sabato, and Naftali Tishby, “Learning and generalization with the information bottleneck,” in International Conference on Algorithmic Learning Theory. Springer, 2008, pp. 92–107.

[28] O. Cowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillelmet, T. Hain, J. Kadlec, V. Karaiskos, et al., “The ani meeting corpus,” in Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research, 2005, vol. 88.