Nonlinear ISA with Auxiliary Variables for Learning Speech Representations

Amrith Setlur†, Barnabás Póczos†, Alan W Black†
†Carnegie Mellon University
asetlur@cs.cmu.edu, bapoczos@cs.cmu.edu, awb@cs.cmu.edu

Abstract

This paper extends recent work on nonlinear Independent Component Analysis (ICA) by introducing a theoretical framework for nonlinear Independent Subspace Analysis (ISA) in the presence of auxiliary variables. Observed high dimensional acoustic features like log Mel spectrograms can be considered as surface level manifestations of nonlinear transformations over individual multivariate sources of information like speaker characteristics, phonological content etc. Under assumptions of energy based models we use the theory of nonlinear ISA to propose an algorithm that learns unsupervised speech representations whose subspaces are independent and potentially highly correlated with the original non-stationary multivariate sources. We show how nonlinear ICA with auxiliary variables can be extended to a generic identifiable model for subspaces as well as also providing sufficient conditions for the identifiability of these high dimensional subspaces. Our proposed methodology is generic and can be integrated with standard unsupervised approaches to learn speech representations with subspaces that can theoretically capture independent higher order speech signals. We evaluate the gains of our algorithm when integrated with the Autoregressive Predictive Coding (APC) model by showing empirical results on the speaker verification and phoneme recognition tasks.

Index Terms: ISA, speech representation learning, unsupervised learning

1. Introduction

The speech signals that we observe can be viewed as high-dimensional surface level manifestations of samples from independent non-stationary sources, that are entangled via a nonlinear mixing mechanism. These sources can be entangled at session, utterance or segment levels [1]. Speech representations learnt by training deep recurrent models [2, 3] over these surface level features fail to capture the original signals in their purest disentangled form. Unsupervised disentanglement of speech representations has been an active area of research [4, 5] since it has been shown that recovering independent factors of variation can improve the performance of downstream tasks like Automatic Speech Recognition (ASR), especially under low resource constraints and domain mismatch [1]. Inspired by this, we propose an algorithm to learn unsupervised speech representations with independent subspaces, each of which can capture distinct disentangled source signals. These distinct subspaces can be potentially informative patterns based on speaker characteristics or suphnonformative events. This can be useful in learning a variety of acoustic models given very few labeled samples for each.

Recently [6] it has been shown that learning disentangled representations is impossible without explicit bias on the algorithm and the data. Hence, we leverage a more principled approach to capturing the independent sources through the lens of nonlinear Independent Subspace Analysis (ISA) in the presence of auxiliary variables.

Nonlinear Independent Component Analysis (ICA) is a provably unidentifiable problem [7] as opposed to linear ICA [8] which is identifiable given non-gaussian sources and other fundamental restrictions on the mixing matrix [9]. Attempts [10, 11, 12] have been made to solve nonlinear ICA for i.i.d distributions under stricter assumptions on the generative process [13, 14]. Recent progress in the field [15, 16] has revolved around a generic identifiable model that renders the latent sources conditionally independent in the presence of auxiliary variables. But most of the work [17, 18] has been focused on unsupervised speech representations with subspaces that can theoretically capture independent higher order speech signals. We extend the auxiliary variable model proposed by [17] for multivariate sources by first stating sufficient conditions for the separability of sources and then, providing training objectives suitable for learning speech representations with finite audio samples. Nonlinear ISA is leveraged to learn unsupervised features on large unlabeled speech datasets. Using these features, simpler (linear) models are learnt on small labeled datasets.

Numerous approaches [5, 19, 20, 21] have been proposed for learning unsupervised speech representations. Recent ones [2, 5] have been based on predictive coding schemes that use language model like objectives. In parallel, there have been efforts to learn quantized representations via temporal segmentation and phonetic clustering [22] so as to map frame representations to linguistic units. But such models are fairly complicated and tricky to train. Also, most of these methods learn highly entangled representations that suffer from spurious correlations in the underlying data and thus fail to generalize. Our proposed algorithm improves upon these approaches by advocating for independent subspaces attained via additional constraints in the original optimization objectives. We begin by providing a theoretically identifiable model for nonlinear ISA and then discuss how the model can be incorporated into existing methods for learning unsupervised speech representations.

2. Theory

We introduce a generative model of the observed data that we assume henceforth and present conditions under which, the original multi-dimensional sources are identifiable. We assume that the observed data \( x \in X \subset \mathbb{R}^{md} \) is generated by applying a non-linear invertible transform \( f \) on \( n \) source signals \( s_1, \ldots, s_n \in S \subset \mathbb{R}^d \). We are given a dataset \( D = \{(x^{(i)}, u^{(i)})\}_{i=1}^N \) with \( N \) samples where each \( x^{(i)} = f(s^{(i)}) \), \( s^{(i)} = \bigotimes_{j=1}^{S} s_j^{(i)} = [s_1^{(i)} \ldots s_n^{(i)}] \). Here, \( u^{(i)} \in U \subset \mathbb{R}^{p} \) denotes the corresponding auxiliary variable for \( x^{(i)} \) and \( f : S^n \rightarrow X \) is a non-linear mixing function (eqn. 1), which is invertible and continuously differentiable almost everywhere (a.e). The objective is to learn representations that can recover the source signals \( \{(s_j^{(i)})_{j=1}^{n}\} \) up to an identifiability factor that we shall define shortly. For

\[ \bigotimes_{j=1}^{S} s_j^{(i)} = [s_1^{(i)} \ldots s_n^{(i)}] \]
notational convenience, we denote the $j^{th}$ scalar element in a vector $z$ as $z_j$ and the $j^{th}$ consecutive $d$-dimensional vector ($j^{th}$ subspace) in $z$ as $z_j$, or as $z_{ij} = [z_{j(1)}d + \ldots + z_{j(d)}]$.

Model The source distributions $\{p_i(s_i)\}_{i=1}^n$ are assumed to be independent given the auxiliary variable $u$ (eqn. 1) and their densities are given by conditional energy based models (eqn. 2) which have universal approximation capabilities [16].

\[
\begin{align*}
\mathbf{x} &= f(s) \\
\log p(s|u) &= \sum_{i=1}^n \log p_i(s_i|u) \\
p_i(s_i|u) &= \exp(\phi_i(s_i)^T \eta_i(u)) / Z_i(u) \\
\phi_i : S \to \mathbb{R}^m \\
\eta_i : \mathcal{U} \to \mathbb{R}^m
\end{align*}
\]

Definition of Identifiability We shall define the original sources $\{s_i\}_{i=1}^n$ to be identifiable if there exists an algorithm that takes as input a pair comprising of the observed sample and the corresponding auxiliary variable ($\mathbf{x} = f(s), u$), and outputs $\{g_i(s_{1n}), \ldots, g_i(s_{in})\}$, for some permutation $\pi : \mathbb{N}^n \to \mathbb{N}^n$ over $\{1 \ldots n\}$. Each $g_i : S \to \mathcal{S}$ is an invertible (a.e) function and is defined as a function of a single distinct source $s_{\pi i}$.

Polynomial algorithms [8] in linear ICA [9] rely on estimators of Mutual Information (MI) to be able to separate the observed mixed samples into samples from the original source signals. Similarly, for nonlinear ICA we compute MI between the observed and auxiliary variables ($I(x, u)$ in eqn. 3) using Noise Contrastive Estimation (NCE) [23]. A nonlinear logistic classifier is used to distinguish between correct (observed) pairs $(x^{(1)}, u^{(1)})$ and randomly generated incorrect pairs $(x^{(1)}, \tilde{u}^{(1)})$. The regression function for this logistic classifier is given by $r(x, u)$ which follows the form in eqn. $26$. Invertible functions can identify the conditionally independent sources $\{s_i\}_{i=1}^n$. For an observed sample $x \in X$, let $y = h^*(x)$ be given by the optimal functions $\{h_i^*\}_{i=1}^n$. The functions $\{h_i, h_i^*\}_{i=1}^n$ are learnt using the NCE objective whose regression function is given by $r(x, u)$. Since $s = f^{-1}(h^{-1}(y))$ is a composition of two invertible transforms, we introduce $v : \mathbb{R}^{2nd} \to \mathbb{S}^n$ where $s = v(y)$. Also, let $f^{-1}u$ be denoted by $g$. From eqn. 3 we know that $r(x, u) = \log p(x|u) - \log p(x)$. Using the density transformation rules [25] for invertible functions we can show that, $\log p(x|u) = \log p(s|u) + \log |\det J_p(x)|$ and $\log p(x) = \log p(s) + \log |\det J(g)(x)|$. Thus, $r(x, u) = \log p(s|u) - \log p(s)$.

Using density transformation rules [25] for invertible functions we can show that, $\log p(x|u) = \log p(s|u) + \log |\det J_p(x)|$ and $\log p(x) = \log p(s) + \log |\det J(g)(x)|$. Thus, $r(x, u) = \log p(s|u) - \log p(s)$.

The subspaces $\{h_i(x)\}_{i=1}^n$ can identify the conditionally independent sources $\{s_i\}_{i=1}^n$ up to the definition of identifiability.

Proof Sketch: For an observed sample $x \in X$, let $y = h^*(x)$ be given by the optimal functions $\{h_i^*\}_{i=1}^n$. The functions $\{h_i, h_i^*\}_{i=1}^n$ are learnt using the NCE objective whose regression function is given by $r(x, u)$. Since $s = f^{-1}(h^{-1}(y))$ is a composition of two invertible transforms, we introduce $v : \mathbb{R}^{2nd} \to \mathbb{S}^n$ where $s = v(y)$. Also, let $f^{-1}u$ be denoted by $g$. From eqn. 3 we know that $r(x, u) = \log p(x|u) - \log p(x)$. Using the density transformation rules [25] for invertible functions we can show that, $\log p(x|u) = \log p(s|u) + \log |\det J_p(x)|$ and $\log p(x) = \log p(s) + \log |\det J(g)(x)|$. Thus, $r(x, u) = \log p(s|u) - \log p(s)$.

For an observed sample $x \in X$, let $y = h^*(x)$ be given by the optimal functions $\{h_i^*\}_{i=1}^n$. The functions $\{h_i, h_i^*\}_{i=1}^n$ are learnt using the NCE objective whose regression function is given by $r(x, u)$. Since $s = f^{-1}(h^{-1}(y))$ is a composition of two invertible transforms, we introduce $v : \mathbb{R}^{2nd} \to \mathbb{S}^n$ where $s = v(y)$. Also, let $f^{-1}u$ be denoted by $g$. From eqn. 3 we know that $r(x, u) = \log p(x|u) - \log p(x)$. Using the density transformation rules [25] for invertible functions we can show that, $\log p(x|u) = \log p(s|u) + \log |\det J_p(x)|$ and $\log p(x) = \log p(s) + \log |\det J(g)(x)|$. Thus, $r(x, u) = \log p(s|u) - \log p(s)$.
can also recover the original multi-dimensional sources. Albeit, NCE algorithm relies on the assumption of infinite samples of positive \((x, u)\) and negative \((x, \bar{u})\) pairs which is rarely true in practice. Hence, along with the NCE objective which learns \(r(x, u)\) that distinguishes between those pairs, we introduce constraints imposed via the HSIC estimator that specifically accounts for independence amongst the subspaces of \(h(x)\). This acts as a strong inductive bias to learn \(\psi^*, h^*\) with finite observed samples of \((x, u)\). HSIC is a kernel based statistical test of independence for two multivariate random variables and is well suited for high dimensional data as opposed to tests \([27, 28, 29]\) based on the power divergence family and characteristic functions which are mainly meant for low-dimensional random variables \([26]\). Given \(\mathcal{D} = \{x^{(i)}, u^{(i)}\}_{i=1}^N\) with \(N\) samples, let the set of features \((h(x^{(i)}))\) be denoted by \(\{y^{(i)} = h(x^{(i)})\}_{i=1}^N\).

For \(\mathbb{R}^d\) dimensional subspaces \(j, k\), let \(y_j \subset \mathbb{R}^d, y_k \subset \mathbb{R}^d\), \(P_y\) denote a Borel probability measure over \(y_j \times y_k\) with \(N\) i.i.d samples \(Z_{jk} := \{y^{(i)}(x^{(i)})\}_{i=1}^N\) drawn from it. If \(\mathcal{F}, \mathcal{G}\) are two Reproducible Kernel Hilbert Spaces (RKHS) equipped with kernels \(k_f, k_h\) then the biased empirical HSIC criterion \(\overline{h}_{jk} \equiv \frac{1}{N} \text{tr}(K_{f}^{(j)}H_k \circ K_{h}^{(k)}H)\) and \(K_f^{(j)}[p, q] = k_f(y^{(p)}, y^{(q)}), K_h^{(k)}[p, q] = k_h(y^{(p)}, y^{(q)}), H = \frac{1}{1+11^2} \in \mathbb{R}^{N \times N}\).

\section{Algorithm (NCE-HSIC)}

We have shown that the NCE algorithm can learn a regression function of the form \(r(x, u)\) (eqn. 3) with optimal predictors \(\psi^*, h^*\) such that the subspaces of \(h^*(x)\) can recover the original sources \(s_i\). Constrained by a finite dataset we use the biased empirical HSIC estimator \(\overline{h}_{jk}\) (lower values imply more independence) as an additional objective while optimizing for \(\psi^*, h^*\). If the true and noisy samples for the NCE algorithm are given by \((x^{(i)}, x^{(i)})\) and \((x^{(i)}, x^{(i)'})\) respectively, then the final loss objective \(\mathcal{L}_{nh}\) for NCE-HSIC is:

\[
\mathcal{L}_{nh} = \frac{1}{N} \sum_{j \in [N]} r(x^{(j)}, u^{(j)'}) - r(x^{(j)}, u^{(j)}) + \lambda \sum_{j, k} \overline{h}_{jk}
\]

\section{Proposed Methodology}

Speech representations that can explicitly capture factors of variation like phoneme identities or speaker traits while being invariant to other factors like underlying pitch contour or background noise \([4, 5]\) have proven to be beneficial since they are less prone to overfitting on spurious correlations in the data. Nevertheless, disentanglement is hard to achieve in general due to the presence of confounding variables \([6]\). In this section, we introduce our approach APC-NCE-HSIC or ANH to learn representations with independent subspaces that can theoretically capture distinct acoustic/linguistic units relevant for downstream tasks like ASR.

Nonlinear ISA provides us with a simple yet principled framework for learning speech representations in the presence of auxiliary variables, which in the case of sequential data like speech can be "time". Learning unsupervised representations can be posed as a problem of recovering from entangled samples the non-stationary sources that are independent given the auxiliary variable (time frame sequence). The NCE-HSIC algorithm can be used to identify original factors of variation via distinct independent subspaces. In order to ensure that the independent subspaces are not only mutually exclusive but are also having a high MI with surface features like Mel-frequency cepstral coefficients (MFCC) or log Mel spectragram (LMS) we build on:

| Auxiliary Variables | The original LMS sequence of length \(T\) is fragmented into time segments \(\{s_{i}\}_{i=1}^{T}\) of length \(\gamma\) and each element \(x_{i}\) in a given segment \(s_{i}\) has its auxiliary variable \(u_{i}, t\) set to the value \(j\), which is nothing but the corresponding segment’s position in the input sequence. The hidden states of the RNN along with the generated auxiliary variables are passed to the NCE module which, first, generates positive \((x_{i}, u_{i})\) and negative \((x_{i}, \bar{u}_{i})\) pairs and then, learns \(\psi^*, \theta^*\) to distinguish between them optimally. Upon the commencement of the unsupervised learning phase, the hidden state for the \(l^{th}\) frame with surface features \(x_{i}\) would comprise of \(n\) subspaces \((h(\theta^*, x_{i}^{(n)})\) that capture different factors of variation, independent for the same value of the auxiliary variable \(u_{i}\). Thus the hidden states can efficiently decouple factors that vary independently locally.

NCE is a powerful tool to predict MI and has been used in recent works like CPC \([3]\) that rely on the NCE objective to distinguish pairs of context vectors from the same or different time segments. This approach is similar to Time Contrastive Learning (TCL) \([7]\) which is an algorithm for nonlinear ICA. Although TCL has only been shown to work for univariate cases and CPC fails to model independent subspaces explicitly, they serve as a strong motivation for our approach which addresses both concerns.

\section{Experiments and Results}

In this section, we empirically evaluate the performance of the proposed ANH algorithm against two baseline models: APC and CPC on two downstream tasks, (1) phoneme recognition (PR) and (2) speaker verification (SV).

\begin{center}
\textbf{Datasets and Implementation} LibriSpeech corpus \([30]\) was used for unsupervised training of the ANH model and other existing approaches based on predictive coding strategies \([19, 3]\). Although our algorithm can be seamlessly integrated into any of these methods, in this work we show empirical results that highlight the performance improvements gained by incorporating the NCE-HSIC criterion into the APC model.
\end{center}
baselines. The datasets for PR and SV were picked from WSJ [31] and TIMIT corpora respectively. For APC we use a multi-layer unidirectional LSTM with residual connections exactly as detailed in [2], with the exception of using 4 layers in the LSTM (wherever mentioned explicitly) and for CPC modifications suggested in [2] are made for a fair comparison. In the unsupervised phase we train the RNN using the $L_{anh}$ objective. The RNN hidden states which are 512-dimensional are assumed to be a collection of $n = 4$ contiguous subspaces each of which has $d = 128$ dimensions. These 4 subspaces of the RNN parameterized by $\theta$, represent the output \{h_i(\theta, x)\}_{i=1}^4$ where $x$ is the LMS feature and $h_i(\theta, x)$ is the $i^{th}$ subspace. The NCE module also needs $\psi(\cdot, \cdot)$ which is implemented using 4-layer MLPs, with ReLU activations, dropsouts and batch-normalization. For $L_{anh}$, five negative pairs are drawn for every positive pair. In the supervised phase, once the ISA features \{h(\theta, x)\} given by the hidden states (final layer) of the trained RNN ($\theta^*$) are extracted, a supervised linear classifier is trained over features from each frame for PR whereas an LDA model is trained over features averaged over the entire sequence for SV.

**Phoneme Recognition** Table 1 highlights the performance (Phone Error Rates (PER)) of our approach (ANH)\(^7\) against the best variants of the CPC, APC models. The supervised baseline (LMS+MLP) which involves training a 3-layer nonlinear classifier over the LMS features fails to capture contextual information. Even though CPC can learn contextual features, it only captures information relevant for recognizing contexts that are $\tau$ steps apart. Thus it may ignore signals that remain relatively stationary for the entire utterance [2]. On the other hand APC directly predicts surface features $\tau$ steps ahead and thus can model sub-phonetic context useful in predicting the next phone. ANH with $\tau = 5$ has the least PER since the addition of the NCE-HSIC objective enables the model to learn noise-free subspaces that can capture relevant factors like formant movements. Finally, adding layers to the RNN further improves the scores.

Table 1: Performance comparison (based on PER) on the Phoneme Recognition task (WSJ corpus [31]).

<table>
<thead>
<tr>
<th>Method</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># looka heed-steps ($\tau$) 2 5 10</td>
</tr>
<tr>
<td>LMS+MLP (supervised)</td>
<td>42.5</td>
</tr>
<tr>
<td>CPC [3]</td>
<td>41.8 44.6 47.3</td>
</tr>
<tr>
<td>APC (3-layer) [2]</td>
<td>36.6 35.7 35.5</td>
</tr>
<tr>
<td>APC (4-layer) [2]</td>
<td>34.5 35.2 33.8</td>
</tr>
<tr>
<td>ANH (3-layer) (Ours)</td>
<td>33.2 31.3 34.7</td>
</tr>
<tr>
<td>ANH (4-layer) (Ours)</td>
<td>31.9 31.8 34.2</td>
</tr>
</tbody>
</table>

**Ablations** NCE-HSIC model when trained without the $L_{apc}$ loss rendered independent subspaces but performed poorly on PR since there is no reason to believe why such subspaces would retain phonetic information. Adding the APC objective aids the model (ANH) to learn acoustic features while disentangling the factors across subspaces (table 1). Removing the HSIC criterion increased the PER and the model training also took \((\times 2)\) longer to converge. This reinforces our hypothesis that the HSIC criterion provides a good inductive bias for a more generalizable model.

**Independence** In order to measure the independence of the four 128-dimensional subspaces of the RNN states, absolute values of the Pearson’s Correlation were computed on the validation splits for PR, SV. When averaged over all possible pairs, they were found to be 0.21, 0.19 on PR, SV respectively when both NCE and HSIC objectives were considered in $L_{anh}$. With $\lambda = 0$ these values were 0.29 and 0.33 but were still significantly lower as compared to the case of APC which had average absolute correlation values of 0.81 and 0.77 on PR and SV respectively.

**Time Segment Length** \((\gamma)\) We show the impact of the time segment length $\gamma$ on the phoneme classification task in table 3. As we increase $\gamma$ the total number of segments (and auxiliary variables) reduce in an utterance. Theoretically, 2nd distinct auxiliary variables are needed to identify $n$ sources each of which is $d$-dimensional (sec. 2). Hence increasing $\gamma$ to values greater than 50 leads to higher ($\geq$ 40) PERs. Additionally, we observe that when the RNN is trained with higher values of $\tau$ for the APC objective PER drops by using wider segments. This may indicate that the distribution of the underlying factors remain stationary for longer periods at higher values of $\tau$.

Table 3: Comparing different values of $\gamma$ for ANH (3-layer) model on the phoneme classification task.

<table>
<thead>
<tr>
<th>Segment size $\gamma$</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td># looka heed-steps ($\tau$) 2 5 10</td>
<td></td>
</tr>
<tr>
<td>$\gamma = 10$</td>
<td>39.4 38.5 36.8</td>
</tr>
<tr>
<td>$\gamma = 20$</td>
<td>38.1 35.3 37.5</td>
</tr>
<tr>
<td>$\gamma = 30$</td>
<td>33.2 31.3 34.7</td>
</tr>
<tr>
<td>$\gamma = 50$</td>
<td>34.0 32.0 33.5</td>
</tr>
</tbody>
</table>

**5. Conclusion**

We extend nonlinear ICA and show how the proposed algorithm to compute MI between the observed and auxiliary variables can provably identify independent subspaces under certain regularity conditions. We also use the algorithm to learn unsupervised speech representations with disentangled subspaces when integrated with existing approaches like APC. Future work may involve a close analysis of the features in these subspaces to understand which orthogonal components are represented by each and how they can prove to be useful for downstream tasks.

\(^7\)For brevity we skip the details of the dataset and refer the reader to [2] from where we borrowed the dataset splits and input LMS features.\(^8\)The optimal $\beta$ in $L_{anh}$ and $\lambda$ in $L_{apc}$ were found to be 0.1 and 0.02.\(^9\)Unless specified all ANH models are trained with $\gamma = 30$.\(^{10}\)ANH model has more than 3 layers. Lowering $\tau$ reduced EER in most cases and had minimal impact on the independence ($\mathbb{E}_{\theta} \lambda_i$).
6. References


