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 while 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 of patterns based on speaker characteristics or subphonetic 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 slightly stronger 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 univariate sources which means that these models can't be directly applied to speech where the source signals are very high dimensional. Hence, we extend the auxiliary variables 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 $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{nd}$ is generated by applying a non-linear invertible transform f on n source signals $\mathbf{s}_1 \ldots \mathbf{s}_n \in \mathcal{S} \subset \mathbb{R}^d$. We are given a dataset $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{u}^{(i)})\}_{i=1}^{N}$ with N samples where each $\mathbf{x}^{(i)} = f(\mathbf{s}^{(i)})$, $\mathbf{s}^{(i)} = \bigotimes_{j=1}^{n} \mathbf{s}_j^{(i)} = [\mathbf{s}_1^{(i)} \ldots \mathbf{s}_n^{(i)}]$[1]. Here, $\mathbf{u}^{(i)} \in \mathcal{U} \subset \mathbb{R}^p$ denotes the corresponding auxiliary variable for $\mathbf{x}^{(i)}$ and $f : \mathcal{S}^n \rightarrow \mathcal{X}$ 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 $(\{\mathbf{s}_i\}_{i=1}^{n})$ up to an identifiability factor that we shall define shortly. For

[1] Here $\bigotimes$ denotes the concatenation operation.

notational convenience, we denote the j^{th} scalar element in a vector $\mathbf{z}$ as z_j and the i^{th} consecutive d-dimensional vector (i^{th} subspace) in $\mathbf{z}$ as $\mathbf{z}_i$ or as $\mathbf{z}_{i:} = \left[z_{(i-1)d+1} \dots z_{id}\right]$.

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

$$\mathbf{x} = f(\mathbf{s}) \qquad \log p(\mathbf{s}|\mathbf{u}) = \sum_{i=1}^{n} \log p_i(\mathbf{s}_i|\mathbf{u}) \tag{1}$$

$$p_i(\mathbf{s}_i|\mathbf{u}) = \frac{\exp \phi_i(\mathbf{s}_i)^T \eta_i(\mathbf{u})}{Z_i(\mathbf{u})} \qquad \begin{matrix} \phi_i : \mathcal{S} \to \mathbb{R}^m \\ \eta_i : \mathcal{U} \to \mathbb{R}^m \end{matrix} \tag{2}$$

Definition of Identifiability We shall define the original sources $\{\mathbf{s}_i\}_{i=1}^n$ to be identifiable if there exists an algorithm that takes as input a pair comprising of the observed sample and the corresponding auxiliary variable ($\mathbf{x} = f(\mathbf{s}), \mathbf{u}$), and outputs $\left[g_1(\mathbf{s}_{\pi_1}), \dots, g_n(\mathbf{s}_{\pi_n})\right]$, for some permutation $\pi : \mathbb{N}^n \to \mathbb{N}^n$ over $\{1 \dots n\}$. Each $g_i : \mathcal{S} \to \mathcal{S}$ is an invertible (*a.e*) function and is defined as a function of a single distinct source $\mathbf{s}_{\pi_i}$.

Popular 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 ($\mathbf{I}(\mathbf{x}, \mathbf{u})$ in *eqn.* 3) using Noise Contrastive Estimation (NCE) [23]. A nonlinear logistic classifier is used to distinguish between correct (observed) pairs $(\mathbf{x}^{(i)}, \mathbf{u}^{(i)})$ and randomly generated incorrect pairs $(\mathbf{x}^{(i)}, \tilde{\mathbf{u}}^{(i)})$ where $\tilde{\mathbf{u}}^{(i)}$ is drawn from the marginal distribution over $\mathbf{u}$. The regression function for this logistic classifier is given by $r(\mathbf{x}, \mathbf{u})$, where $h_i : \mathcal{X} \to \mathbb{R}^d$, $\psi_i : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R} \in L^2$ are sufficiently smooth universal function approximators (neural networks) and $\forall i$, h_i is invertible *a.e.*

$$\mathbf{I}(\mathbf{x}, \mathbf{u}) = \int_{x,u} \log \frac{p(\mathbf{x}, \mathbf{u})}{p(\mathbf{x})p(\mathbf{u})} \, d\mathbb{P}(\mathbf{x}, \mathbf{u}) = \int r(\mathbf{x}, \mathbf{u}) \, d\mathbb{P}(\mathbf{x}, \mathbf{u})$$

$$\textit{where,} \quad r(\mathbf{x}, \mathbf{u}) = \sum_{i=1}^{n} \psi_i(h_i(\mathbf{x}), \mathbf{u}) \tag{3}$$

The following main ISA separation theorem states that the vector $h(\mathbf{x}) = \bigotimes_{i=1}^n h_i(\mathbf{x}) \in \mathbb{R}^{nd}$, with subspaces $h_i(\mathbf{x}) \in \mathbb{R}^d$ can recover $\mathbf{s}_i$ since $\exists \pi, \{g_i\}_{i=1}^n$ such that $h_i(\mathbf{x}) = g_i(\mathbf{s}_{\pi_i})$.

Theorem 1. *Given that we observe the dataset $\mathcal{D}$ with N samples: $\{\mathbf{x}^{(i)} = f(\mathbf{s}^{(i)}), \mathbf{u}^{(i)}\}_{i=1}^N$ generated by a model based on eqns. (1, 2), then under the following assumptions*[2]*:*

1. ***Realizability Assumption:*** *Given infinite ($N \to \infty$) samples one can efficiently learn ψ_i^*, h_i^* such that the NCE algorithm can estimate the mutual information $\mathbf{I}(\mathbf{x}, \mathbf{u})$ with an arbitrarily small error, using the regression function $r(\mathbf{x}, \mathbf{u})$ which follows the form in eqn. 3.*
2. ***Separability Assumption:*** *$\forall \mathbf{s} \in \mathcal{S}^n$, $\mathbf{z} \neq 0 \in \mathbb{R}^d$ with first and second order derivatives given by tensors $\nabla\phi_i(\mathbf{s}_i) \in \mathbb{R}^{m \times d}$ and $\nabla^2\ \phi_i(\mathbf{s}_i) \in \mathbb{R}^{m \times d \times d}$ respectively; $\exists \{\mathbf{u}_l\}_{l=0}^{2nd} \in \mathcal{U}^{2nd+1}$ such that:*

$$\left\{ \bigotimes_{i=1}^{n} \left(\begin{bmatrix} \nabla\phi_i(\mathbf{s}_i)^T \\ (\nabla^2\phi_i(\mathbf{s}_i) \, \bar{\times}_3{}^{3} \, \mathbf{z})^T \end{bmatrix} \zeta_i(\mathbf{u}_l, \mathbf{u}_0) \right) \right\}_{l=1}^{2nd}$$

spans $\mathbb{R}^{2nd}$ for $\zeta_i(\mathbf{u}_l, \mathbf{u}_0) = (\eta_i(\mathbf{u}_l) - \eta_i(\mathbf{u}_0))$, the subspaces $\{h_i(\mathbf{x})\}_{i=1}^n$ can identify the conditionally independent sources $\{\mathbf{s}_i\}_{i=1}^n$ up to the ***definition of identifiability****.*

Proof Sketch[4]*:* For an observed sample $\mathbf{x} \in \mathcal{X}$, let $\mathbf{y} = h^*(\mathbf{x})$ be given by the optimal functions $\{h_i^*\}_{i=1}^n$. The functions $\{\psi_i^*, h_i^*\}_{i=1}^n$[5] are learnt using the NCE objective whose regression function is given by $r(\mathbf{x}, \mathbf{u})$. Since $\mathbf{s} = f^{-1}(h^{-1}(\mathbf{y}))$ is a composition of two invertible transforms, we introduce $v : \mathbb{R}^{nd} \to \mathcal{S}^n$ where $\mathbf{s} = v(\mathbf{y})$. Also, let f^{-1} be denoted by g. From *eqn.* 3 we know that $r(\mathbf{x}, \mathbf{u}) = \log p(\mathbf{x}|\mathbf{u}) - \log p(\mathbf{x})$,

Using the density transformation rules [25] for invertible functions we can show that, $\log p(\mathbf{x}|\mathbf{u}) = \log p(\mathbf{s}|\mathbf{u}) + \log |\det \mathbf{J}g(\mathbf{x})|$ and $\log p(\mathbf{x}) = \log p(\mathbf{s}) + \log |\det \mathbf{J}g(\mathbf{x})|$. Thus, $r(\mathbf{x}, \mathbf{u}) = \log p(\mathbf{s}|\mathbf{u}) - \log p(\mathbf{s})$. Using *eqn.* 3:

$$\sum_{i=1}^{n} \psi_i^*(\mathbf{y}_i, \mathbf{u}) = \log p(v(\mathbf{y})|\mathbf{u}) - \log p(v(\mathbf{y})) \tag{4}$$

We begin by substituting *eqns.* 1, 2 in the above result. Also, since *eqn.* 4 holds true for $\{\mathbf{u}_l\}_{l=0}^{2nd}$, we can get $2nd + 1$ such equations and from each we can subtract the equation given by $\mathbf{u}_0$, which leaves us with $2nd$ *eqns.* of the form $\sum_{i=1}^n \phi_i(v(\mathbf{y})_{i:})^T \zeta_i(\mathbf{u}_l, \mathbf{u}_0) - (\log Z_i(\mathbf{u}_l) - \log Z_i(\mathbf{u}_0)) = \sum_{i=1}^n \psi_i^*(\mathbf{y}_i, \mathbf{u})$. Taking the derivative of both sides of this *eqn.* *w.r.t.* y_j and subsequently w.r.t y_k *s.t.* $\lceil j/d \rceil \neq \lceil k/d \rceil$ we get,

$$0 = \sum_{i=1}^{n} \left(\underbrace{\nabla\phi_i(v(\mathbf{y})_{i:})}_{①} \frac{\partial^2 v(\mathbf{y})_{i:}}{\partial y_j \partial y_k} \right)^T \zeta_i(\mathbf{u}_l, \mathbf{u}_0)$$
$$+ \left(\underbrace{\left(\nabla^2\phi_i(v(\mathbf{y})_{i:}) \bar{\times}_3 \frac{\partial v(\mathbf{y})_{i:}}{\partial y_j} \right)}_{②} \frac{\partial v(\mathbf{y})_{i:}}{\partial y_k} \right)^T \zeta_i(\mathbf{u}_l, \mathbf{u}_0)$$

Concatenating ①, ② into a single matrix in $\mathbb{R}^{2d \times m}$, the above can be written as a single euclidean inner product in $\mathbb{R}^{2nd}$.

$$\left(\bigotimes_{i=1}^{n} \left(\begin{bmatrix} \nabla\phi_i(\mathbf{s}_i)^T \\ \left(\nabla^2\phi_i(\mathbf{s}_i) \, \bar{\times}_3 \, \frac{\partial v(\mathbf{y})_{i:}}{\partial y_j}\right)^T \end{bmatrix} \zeta_i(\mathbf{u}_l, \mathbf{u}_0) \right) \right) \Gamma(\mathbf{y}) = 0$$

For $\Gamma(\mathbf{y}) = \left(\bigotimes_{i=1}^{n} \left[\frac{\partial^2 v(\mathbf{y})_{i:}}{\partial y_j \partial y_k} \; \frac{\partial v(\mathbf{y})_{i:}}{\partial y_k} \right] \right) \in \mathbb{R}^{2nd}$ the above equation holds true for $2nd$ distinct values of the auxiliary variable $\mathbf{u}_l$. For invertible v, if we assume that $\frac{\partial v(\mathbf{y})_{i:}}{\partial y_j} \neq 0$ then we can apply the **separability** *assm.* which implies $\Gamma(\mathbf{y}) = \mathbf{0}$. This further implies that $\frac{\partial v(\mathbf{y})_{i:}}{\partial y_k} = 0$. Thus $\forall i$, $\frac{\partial v(\mathbf{y})_{i:}}{\partial y_j} \vee \frac{\partial v(\mathbf{y})_{i:}}{\partial y_k}$. Since $\lceil j/d \rceil \neq \lceil k/d \rceil$, y_j and y_k belong to distinct subspaces of $y = h(\mathbf{x})$. Hence the i^{th} source given by $v(\mathbf{y})_{i:}$ cannot simultaneously be a function of two distinct subspaces of $h(\mathbf{x})$. Given the invertible function $f(h(\cdot))$ with its full rank jacobian we can recover the sources $\{\mathbf{s}_i\}_{i=1}^n$ via the subspaces of $\mathbf{h}(\mathbf{x})$; $h_i(\mathbf{x}) = g_i(\mathbf{s}_{\pi_i})$ for an invertible function g_i, permutation π.

Hilbert-Schmidt Independence Criterion (HSIC) [26] The above theorem proves the existence of functions ψ^*, h^* that can not only compute $\mathbf{I}(\mathbf{x}, \mathbf{u})$ with arbitrary precision but

[2] The **separability** *assm.* requires the auxiliary variables $\mathbf{u}$ to have a sufficiently strong and diverse effect on the source distributions [17].

[3] $\bar{\times}_3$ denotes the 3^{rd} mode product [24].

[4] For more details on the validity and necessity of similar results for independent components (as opposed to subspaces) we refer the reader to [17]. Also, for the sake of completion we show a proof sketch for the identifiability of our ISA model. It's an extension of the proof for the univariate case [17, 18].

[5] Subscript i is dropped wherever it can be understood from context.

can also recover the original multi-dimensional sources. Albeit, NCE algorithm relies on the assumption of infinite samples of positive $(\mathbf{x}, \mathbf{u})$ and negative $(\mathbf{x}, \tilde{\mathbf{u}})$ pairs which is rarely true in practice. Hence, along with the NCE objective which learns $r(\mathbf{x}, \mathbf{u})$ that distinguishes between those pairs, we introduce constraints imposed via the HSIC estimator that specifically accounts for independence amongst the subspaces of $h(\mathbf{x})$. This acts as a strong inductive bias to learn ψ^*, h^* with finite observed samples of $(\mathbf{x}, \mathbf{u})$. 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 $\mathcal{D} = \{\mathbf{x}^{(i)}, \mathbf{u}^{(i)}\}_{i=1}^{N}$ with N samples, let the set of features $(h(\mathbf{x}^{(i)}))$ be denoted by $\{\mathbf{y}^{(i)} = h(\mathbf{x}^{(i)})\}_{i=1}^{N}$. For $\mathbb{R}^d$ dimensional subspaces j, k let $\mathbf{y}_j \in \mathcal{Y}_j \subseteq \mathbb{R}^d$, $\mathbf{y}_k \in \mathcal{Y}_k \subseteq \mathbb{R}^d$ and $\mathbf{P}_{jk}$ denote a Borel probability measure over $\mathcal{Y}_j \times \mathcal{Y}_k$ with N i.i.d samples $\mathcal{Z}_{jk} := \{(\mathbf{y}_j^{(i)}, \mathbf{y}_k^{(i)})\}_{i=1}^{N}$ drawn from it. If $\mathcal{F}, \mathcal{G}$ are two Reproducible Kernel Hilbert Spaces (RKHS) equipped with kernels[6] k_f, k_g then the biased empirical HSIC criterion $\hat{\mathbb{H}}_{jk} = \frac{1}{N^2}\mathrm{tr}(\mathbf{K_f}^{(j)}\mathbf{H}\mathbf{K_g}^{(k)}\mathbf{H})$ and $\mathbf{K_f}^{(j)}[p,q] = k_f(\mathbf{y}_j^{(p)}, \mathbf{y}_j^{(q)})$, $\mathbf{K_g}^{(k)}[p,q] = k_g(\mathbf{y}_k^{(p)}, \mathbf{y}_k^{(q)})$, $\mathbf{H} = \mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^T \in \mathbb{R}^{N \times N}$.

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

$$\mathcal{L}_{nh} = \frac{1}{N}\sum_{l \in [N]} r(\mathbf{x}^{(l)}, \mathbf{u}^{(l' \neq l)}) - r(\mathbf{x}^{(l)}, \mathbf{u}^{(l)}) + \lambda \sum_{j,k} \hat{\mathbb{H}}_{jk}$$

3. 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 spectrograms (LMS) we build on 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.

APC (Autoregressive Predictive Coding) [2] is a language model based method to learn unsupervised speech representations. It uses an RNN to model temporal information within an acoustic sequence comprising of 80-dimensional LMS features $\{\mathbf{x}_i\}_{i=0}^{T}$. Given these features until a fixed time step t, the APC model predicts the surface feature τ time steps ahead *i.e.* $\mathbf{x}_{t+\tau}$. If $\{\hat{\mathbf{p}}_i\}_{i=0}^{T-\tau}$ represents the sequence predicted by the RNN, then the l_1 loss used to train the model is given by:

$$\mathcal{L}_{apc}(\mathbf{x}) = \sum_{i=0}^{T-\tau} - \log p(\mathbf{x}_{i+\tau}|\mathbf{x}_1 \ldots \mathbf{x}_i) = \sum_{i=0}^{T-\tau} |\hat{\mathbf{p}}_i - \mathbf{x}_{i+\tau}|$$

APC-NCE-HSIC or ANH is our proposed model where features with independent subspaces are learnt through the NCE-HSIC criterion which is applied to the hidden states of the RNN module trained with the APC objective above. Specifically, the function $h(\mathbf{x})$ is modeled using the RNN. The NCE-HSIC criterion increases the correlation of the original sources with the subspaces of $h(\mathbf{x})$ or in this case the subspaces of the hidden states of the RNN. If the RNN is parameterized by $\theta \in \Theta$ then the hidden state can be represented as the function $h(\theta, \mathbf{x})$. With $r(\mathbf{x}, \mathbf{u}) = \sum_{i=1}^{n} \psi_i(h_i(\theta, \mathbf{x}), \mathbf{u})$ the final objective would be:

$$\underset{\{\psi_i\}_{i=1}^{n}, \theta}{\operatorname{argmin}} \; \mathcal{L}_{anh} = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} \mathcal{L}_{apc}(\mathbf{x}) + \beta \mathcal{L}_{nh} \tag{5}$$

Auxiliary Variables[7] The original LMS sequence of length T is fragmented into time segments $\{s_j\}_{j=1}^{\lceil T/\gamma \rceil}$ of length γ, and each element $\mathbf{x}_{j,t}$ in a given segment s_j has its auxiliary variable $\mathbf{u}_{j,t}$ 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 $(\mathbf{x}_t, \mathbf{u}_t)$ and negative $(\mathbf{x}_t, \tilde{\mathbf{u}}_t)$ pairs and *then*, learns ψ^*, θ^* to distinguish between them optimally. Upon the commencement of the unsupervised learning phase, the hidden state for the t^{th} frame with surface features $\mathbf{x}_t$ would comprise of n subspaces $(\{h_i(\theta^*, \mathbf{x}_t)\}_{i=1}^{n})$ that capture different factors of variation, independent for the same value of the auxiliary variable $\mathbf{u}_t$. 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** [8] [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.

4. 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).

Datasets and Implementation LibriSpeech corpus [30] was used for unsupervised training of the ANH model and other

[6] $k_f : \mathcal{Y}_j \times \mathcal{Y}_j \rightarrow \mathbb{R}$, $k_g : \mathcal{Y}_k \times \mathcal{Y}_k \rightarrow \mathbb{R}$; for $\mathbf{z}, \mathbf{z}' \in \mathcal{Y}_j$, $k_f(\mathbf{z}, \mathbf{z}') = \langle k_f(\mathbf{z}, \cdot), k_f(\mathbf{z}', \cdot)\rangle_{\mathcal{F}}$ and for $\mathbf{z}, \mathbf{z}' \in \mathcal{Y}_k$, $k_g(\mathbf{z}, \mathbf{z}') = \langle k_g(\mathbf{z}, \cdot), k_g(\mathbf{z}', \cdot)\rangle_{\mathcal{G}}$.

[7] Auxiliary variables can be potentially given by other domains like the frequency spectrum, but in this work we focus only on time.

[8] Contrastive Predictive Coding.

baselines. The datasets for PR and SV were picked from WSJ [31] and TIMIT corpora respectively [9]. For APC we use a multilayer 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 $\mathcal{L}_{anh}$ 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 θ, represent the output $\{h_i(\theta, \mathbf{x})\}_{i=1}^{4}$ where $\mathbf{x}$ is the LMS feature and $h_i(\theta, \mathbf{x})$ is the i^{th} subspace. The NCE module also needs $\psi_i(\cdot, \cdot)$ which is implemented using 4-layer MLPs, with ReLU activations, dropouts and batch-normalization. For $\mathcal{L}_{nh}$[10], five negative pairs are drawn for every positive pair. In the *supervised* phase, once the ISA features $(h(\theta^*, \mathbf{x}))$ given by the hidden states (final layer) of the trained RNN (θ^*) 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)[11] 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 τ 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 τ steps ahead and thus can model subphonetic 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]).*

Method	**PER**		
# lookahead-steps (τ)	2	5	10
LMS+MLP (supervised)		42.5	
CPC [3]	41.8	44.6	47.3
APC (3-layer) [2]	36.6	35.7	35.5
APC (4-layer) [2]	34.5	35.2	**33.8**
ANH (3-layer) (Ours)	33.2	**31.3**	34.7
ANH (4-layer) (Ours)	**31.9**	31.8	34.2
Ablations	2	5	10
APC + NCE	32.0	32.4	34.3
NCE-HSIC	49.8	48.5	54.6
NCE	49.4	53.2	55.9

Speaker Verification Results for SV are summarized in table 2 which shows lower Equal Error Rates (EER) achieved by ANH as compared to the baselines. It has been shown that in deep language models, lower layers model local syntax while the higher ones capture semantic content [2, 32]. We make similar observations since the EER values increase (for all τ) when the ANH model has more than 3 layers. Lowering τ reduced EER in most cases and had minimal impact on the independence $(\hat{\mathbb{H}}_{jk})$.

Table 2: *Performance (based on EER) on the speaker verification task (*TIMIT *corpus). (*choosing different layers [2])*

Method	**EER**			
# lookahead-steps (τ)	2	3	5	10
CPC features [3]	5.62	5.29	5.42	6.01
APC (3-layer)-1* [2]	3.82	3.67	3.88	4.01
APC (3-layer)-2* [2]	**3.41**	3.72	3.92	4.04
ANH (2-layer) (Ours)	3.53	3.35	3.91	4.12
ANH (3-layer) (Ours)	3.45	**3.12**	**3.45**	**3.67**

Ablations NCE-HSIC model when trained without the $\mathcal{L}_{apc}$ 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 (×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 $\mathcal{L}_{nh}$. 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 (γ) We show the impact of the time segment length γ on the phoneme classification task in table 3. As we increase γ 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 γ to values greater than 50 leads to higher (> 40) PERs. Additionally, we observe that when the RNN is trained with higher values of τ 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 τ.

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

Segment size γ	**PER**		
# lookahead-steps (τ)	2	5	10
$\gamma = 10$	39.4	38.5	36.8
$\gamma = 20$	38.1	35.3	37.5
$\gamma = 30$	**33.2**	**31.3**	34.7
$\gamma = 50$	34.0	32.0	**33.5**

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.

[9] 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.

[10] The optimal β in $\mathcal{L}_{anh}$ & λ in $\mathcal{L}_{nh}$ were found to be 0.1 and 0.02.

[11] Unless specified all ANH models are trained with $\gamma = 30$.

6. References

[1] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Advances in neural information processing systems*, 2017, pp. 1878–1889.

[2] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," *arXiv preprint arXiv:1904.03240*, 2019.

[3] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[4] Y. Li and S. Mandt, "Disentangled sequential autoencoder," *arXiv preprint arXiv:1803.02991*, 2018.

[5] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2041–2053, 2019.

[6] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," *arXiv preprint arXiv:1811.12359*, 2018.

[7] A. Hyvärinen and P. Pajunen, "Nonlinear independent component analysis: Existence and uniqueness results," *Neural Networks*, vol. 12, no. 3, pp. 429–439, 1999.

[8] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[9] A. Hyvarinen, "Survey on independent component analysis," *Neural computing surveys*, vol. 2, no. 4, pp. 94–128, 1999.

[10] Y. Tan, J. Wang, and J. M. Zurada, "Nonlinear blind source separation using a radial basis function network," *IEEE transactions on neural networks*, vol. 12, no. 1, pp. 124–134, 2001.

[11] L. B. Almeida, "Misep–linear and nonlinear ica based on mutual information," *Journal of Machine Learning Research*, vol. 4, no. Dec, pp. 1297–1318, 2003.

[12] L. Dinh, D. Krueger, and Y. Bengio, "Nice: Non-linear independent components estimation," *arXiv preprint arXiv:1410.8516*, 2014.

[13] P. Brakel and Y. Bengio, "Learning independent features with adversarial nets for non-linear ica," *arXiv preprint arXiv:1710.05050*, 2017.

[14] J. A. Lee, C. Jutten, and M. Verleysen, "Non-linear ica by using isometric dimensionality reduction," in *International Conference on Independent Component Analysis and Signal Separation*. Springer, 2004, pp. 710–717.

[15] I. Khemakhem, D. P. Kingma, and A. Hyvärinen, "Variational autoencoders and nonlinear ica: A unifying framework," *arXiv preprint arXiv:1907.04809*, 2019.

[16] I. Khemakhem, R. P. Monti, D. P. Kingma, and A. Hyvärinen, "Icebeem: Identifiable conditional energy-based deep models," *arXiv preprint arXiv:2002.11537*, 2020.

[17] A. Hyvarinen, H. Sasaki, and R. E. Turner, "Nonlinear ica using auxiliary variables and generalized contrastive learning," *arXiv preprint arXiv:1805.08651*, 2018.

[18] A. Hyvarinen and H. Morioka, "Unsupervised feature extraction by time-contrastive learning and nonlinear ica," in *Advances in Neural Information Processing Systems*, 2016, pp. 3765–3773.

[19] Y.-A. Chung and J. Glass, "Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech," *arXiv preprint arXiv:1803.08976*, 2018.

[20] B. Milde and C. Biemann, "Unspeech: Unsupervised speech context embeddings," *arXiv preprint arXiv:1804.06775*, 2018.

[21] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," *arXiv preprint arXiv:1704.04222*, 2017.

[22] A. H. Liu, T. Tu, H.-y. Lee, and L.-s. Lee, "Towards unsupervised speech recognition and synthesis with quantized speech representation learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7259–7263.

[23] M. U. Gutmann and A. Hyvärinen, "Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics," *Journal of Machine Learning Research*, vol. 13, no. Feb, pp. 307–361, 2012.

[24] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[25] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[26] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola, "A kernel statistical test of independence," in *Advances in neural information processing systems*, 2008, pp. 585–592.

[27] T. R. Read and N. A. Cressie, *Goodness-of-fit statistics for discrete multivariate data*. Springer Science & Business Media, 2012.

[28] A. Kankainen, *Consistent testing of total independence based on the empirical characteristic function*. University of Jyväskylä, 1995, vol. 29.

[29] A. Feuerverger, "A consistent test for bivariate dependence," *International Statistical Review/Revue Internationale de Statistique*, pp. 419–433, 1993.

[30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[31] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[32] M. E. Peters, M. Neumann, L. Zettlemoyer, and W.-t. Yih, "Dissecting contextual word embeddings: Architecture and representation," *arXiv preprint arXiv:1808.08949*, 2018.