Towards Robust Waveform-Based Acoustic Models

Dino Oglic, Zoran Cvetkovic, Peter Sollich, Steve Renals, and Bin Yu

***Abstract*—We study the problem of learning robust acoustic models in adverse environments, characterized by a significant mismatch between training and test conditions. This problem is of paramount importance for the deployment of speech recognition systems that need to perform well in unseen environments. First, we characterize data augmentation theoretically as an instance of vicinal risk minimization, which aims at improving risk estimates during training by replacing the delta functions that define the empirical density over the input space with an approximation of the marginal population density in the vicinity of the training samples. More specifically, we assume that local neighborhoods centered at training samples can be approximated using a mixture of Gaussians, and demonstrate theoretically that this can incorporate robust inductive bias into the learning process. We then specify the individual mixture components implicitly via data augmentation schemes, designed to address common sources of spurious correlations in acoustic models. To avoid potential confounding effects on robustness due to information loss, which has been associated with standard feature extraction techniques (e.g., FBANK and MFCC features), we focus on the waveform-based setting. Our empirical results show that the approach can generalize to unseen noise conditions, with 150% relative improvement in out-of-distribution generalization compared to training using the standard risk minimization principle. Moreover, the results demonstrate competitive performance relative to models learned using a training sample designed to match the acoustic conditions characteristic of test utterances.**

***Index Terms*—vicinal risk minimization, out-of-distribution generalization, data augmentation, waveform-based models.**

I. Introduction

We consider the problem of improving the performance of acoustic models in adverse environments, where there is a significant mismatch between training and testing acoustic conditions. This problem is of paramount importance for the deployment of automatic speech recognition systems that are required to perform well in unseen environments, without suffering any significant performance degradation. Recently, there has been considerable progress in improving the performance of acoustic models in adverse conditions for filterbank features [e.g., see 1–4]. However, there are still significant performance gaps for testing environments characterized as novel relative to the ones seen during training. To the best of our knowledge, there has not been a study on the impact of adverse environments on waveform-based acoustic models [5–7], typically encountered in truly end-to-end speech recognition systems. The waveform-based setting is particularly interesting because it allows for an empirical evaluation free of confounding effects on robustness due to information loss and non-adaptive feature extraction process. More specifically, several comparative studies of automatic and human speech recognition [8–10] suggest that the information loss inherent to filterbank features can adversely affect robustness to standard environmental distortions arising from additive and channel (linear filtering) noise [11, 12]. The main challenge comes from the high variability of speech signals that are representative of a sub-phonetic unit. This can be caused by differences between speakers (e.g., accents, pronunciations, speaking styles, emotional states, etc.), environmental noise, different microphones, reverberation, and recording devices [1]. Moreover, the sources of variability are typically non-stationary and can interact with speech signals in a non-linear way [13]. Hence, it is difficult if not impossible to avoid a mismatch between training and testing environments.

At the core of modern automatic speech recognition systems are deep learning models that exploit associations between frame representations and corresponding sub-phonetic units when learning to generalize from training data to the population level. A problem arises when such associations are characteristic of the training sample but are not present in the test samples or, in general, at the population level. This phenomenon is known as spurious correlation and it hinders the generalization abilities of acoustic models. Typically, spurious correlations are scrambled by augmenting the training samples with acoustic conditions that resemble those expected within testing environments. In this regard, a particularly influential learning regime is multi-condition/style training where inputs are transformed by naturally occurring additive noise signals with various signal-to-noise ratios [2, 4, 14]. The noise types are usually selected such that they reduce the mismatch between training and testing conditions; there are several publicly available databases of naturally occurring environment noise signals. This type of data augmentation is practical because it preserves sub-phonetic labels assigned to frames of the original speech signal (i.e., there is no need for re-alignment of augmented utterances). While multi-condition training allows for significant performance improvement in approximately matched testing conditions, the error rates are severely increased in environments with novel speech variability characteristics.

This work was supported in part by EPSRC under Grant EP/R012067/1.

D. Oglic is with the Applied Analytics and AI, Data Sciences and AI, BioPharmaceuticals R&D, AstraZeneca, Cambridge CB2 8PA, UK. This work was done in part while he was with the Department of Engineering, King's College London, London WC2R 2LS, UK. Correspondence to: dino.oglic@astrazeneca.com.

Z. Cvetkovic is with the Department of Engineering, King's College London, London WC2R 2LS, UK (e-mail: zoran.cvetkovic@kcl.ac.uk).

P. Sollich is with the Department of Mathematics, King's College London, London WC2R 2LS, UK, and also the Institute for Theoretical Physics, University of Göttingen, 37073 Göttingen, Germany (e-mail: peter.sollich@kcl.ac.uk).

S. Renals is with the Center for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB, UK (e-mail: s.renals@ed.ac.uk).

B. Yu is with the Departments of Statistics and Electrical Engineering and Computer Sciences, UC Berkeley, Berkeley, CA 94720, USA (e-mail: binyu@stat.berkeley.edu).

Thus, the main shortcoming of such approaches is the fact that it is infeasible to enumerate all possible noise types combined with different signal-to-noise ratios, as well as other sources of variability that one can expect in testing environments [15].

To tackle this issue and allow for learning of robust acoustic models, we propose an approach based on vicinal risk minimization, which aims at improving risk estimates during training by replacing the delta functions that define the empirical density over the input space with an approximation of the population density in the vicinity of the training samples. We assume that the vicinal densities can be described using a mixture of Gaussians and characterize the individual mixture components implicitly via data augmentation schemes that are designed to scramble frequent sources of spurious correlations in speech (Section IV). In our approach, we rely on white noise as a source of randomness and couple it with linear transformations of the input signal that are based on band-pass and band-stop filtering. This purely synthetic approach differs from prior work that typically resorts to databases of standard environment noise types and/or acoustic conditions. The focus of our study is on two types of mismatch between training and test environments. In the first case, we are interested in improving out-of-distribution generalization of waveform-based models when the difference between the two environments lies in the level and type of background noise. The second case, on the other hand, deals with spurious correlations introduced by different microphones used for recording. The latter is challenging to address for waveform-based models due to the fact that the feature extraction process is fully automated and utterance level mean-normalization cannot be performed as in the case of standard non-adaptive filterbank features. Prior work has established that such normalizations can be fundamental in dealing with spurious correlations introduced by different microphones and stationary signal corruptions [2, 16].

Our theoretical contributions include a characterization of robustness relative to the waveform domain, given in terms of the Jacobian and Hessian tensors of the sufficient statistic (Section III). In addition to this, we provide insights into how vicinal risk minimization affects the learning process in terms of inductive bias — data fitting at the level of local neighborhoods centered at training observations instead of pointwise mappings characteristic of standard empirical risk minimization.

The results of our empirical evaluation are presented in Section VI. The main focus of the evaluation is on the generalization from clean to noisy speech, with a severe mismatch between training and run-time conditions. To avoid possible confounding effects [8–10] of the standard feature extraction process on the robustness our empirical study quantifies the effectiveness of the approach relative to the waveform-based models. We conduct our analysis on AURORA4 using the Kaldi *clean-condition recipe* and a recently proposed neural architecture for waveform-based speech recognition called PARZNETS 2D [7]. Our empirical results demonstrate that the proposed approach can improve out-of-distribution generalization abilities of models trained on clean speech by more than 150% in relative terms. Moreover, the results obtained are competitive with the best possible augmentation principle where training samples are transformed using noise types that appear in the test fold. Having demonstrated that the approach can aid generalization to unseen conditions, we evaluate it on conversational speech, thus showing its utility for generalization from headset recorded to distant-talking speech.

II. Parzen Filters and Parznets

In this section, we provide a brief review of Parzen filters and PARZNETS 2D [7] – the waveform based neural architecture used in our empirical evaluations. The proposed data augmentation schemes depend in part on band-pass filtering and for simplicity we have used Parzen filters as a realization of this operator. PARZNETS 2D is a neural architecture that provides an effective means for fully automated feature extraction directly from speech signals. This is achieved by the first layer of the network, which is defined using parametric convolutions that allow for efficient learning of band-pass filters.

In speech recognition, band-pass filtering of signals is traditionally performed by weighted averaging of power spectra, computed over segments of fixed duration. Alternatively, the signal can be convolved with a filter directly in the time-domain. Motivated by this, the first layer of PARZNETS 2D is designed to emulate this operation via a parametric time-domain convolutional operator. To that end, the network employs a family of differentiable band-pass filters based on cosine modulations of compactly supported Parzen windows [17]. In particular, our empirical analysis employs a squared Epanechnikov window function [18]

$$k_\gamma(t) = \begin{cases} \left(1-\gamma t^2\right)^2 & |t| \leq 1/\sqrt{\gamma} \\ 0 & \text{otherwise} \end{cases},$$

where γ is a parameter controlling the window width. To allow for flexible placement of the center/modulation frequency, the filterbank relies on cosine modulation. Thus, Parzen filters are defined with only two differentiable parameters, η controlling the modulation frequency and γ controlling the filter bandwidth,

$$\phi_{\eta,\gamma}(t) = \cos(2\pi\eta t) \cdot k_\gamma(t) \quad . \tag{1}$$

As illustrated in Figure 1, for each filter configuration $\{(\eta_i, \gamma_i)\}_{i=1}^{B}$, Eq. (1) is used to generate a one dimensional filter with maximum length given by the number of samples in 25 ms of speech; filters with shorter support are symmetrically padded with zeros.

PARZNETS 2D take the outputs of parametric convolutions and concatenate them into a spectro-temporal decomposition of a signal, which is then passed to a max pooling operator, followed by layer normalization [19]. The outputs of the parametric convolutional block are then passed to a sequence of standard convolutional operators that perform further band-pass filtering and compression of the signal by different max pooling operators. The convolutional blocks generate a set of *automatically extracted features*, which are then passed to a multi-layer perceptron with four hidden layers.

III. Theory

Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact set containing raw speech frames of fixed duration in its interior (e.g., 200 ms long frames of speech) and $\mathcal{Y}$ the space of categorical labels (e.g., state ids

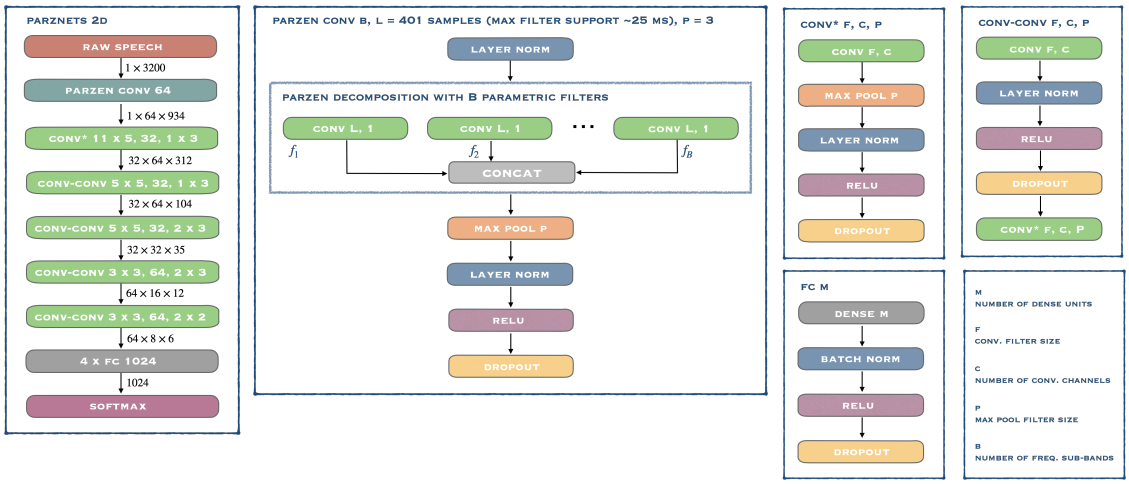

Fig. 1: The figure describes the architecture for PARZNETS with 2D convolutional operators. This is supplemented with an illustration of the Parzen convolutional block that decomposes a raw speech frame into frequency sub-bands.

in hybrid HMM-DNN models). Suppose that a set of labeled examples $\{(x_i, y_i)\}_{i=1}^n$ has been drawn independently from a latent Borel probability measure defined on $\mathcal{X} \times \mathcal{Y}$.

Deep learning models with SOFTMAX outputs (e.g., feedforward neural architectures) typically assume that the conditional probability of a label $y \in \mathcal{Y}$ given an instance $x \in \mathcal{X}$ can be approximated with an exponential family model [20]

$$p\left(y \mid x, \alpha, W\right) = \frac{\exp\left(\alpha_y^\top \Psi\left(x \mid W\right)\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\alpha_{y'}^\top \Psi\left(x \mid W\right)\right)} \ , \qquad (2)$$

where $\alpha \in \mathbb{R}^{D \times |\mathcal{Y}|}$ is a parameter matrix with columns $\alpha_y \in \mathbb{R}^D$ defining the *softmax probabilities* $p\left(y \mid x, \alpha, W\right)$ and $\Psi\left(x \mid W\right) \in \mathbb{R}^D$ is a sufficient statistic of x, given by pre-softmax neural network parameters W. This model can also be written as a special case of a conditional exponential family model [21]

$$p\left(y \mid x, \alpha, W\right) = \frac{\exp\left(\alpha^\top \Phi\left(x, y \mid W\right)\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\alpha^\top \Phi\left(x, y' \mid W\right)\right)} \ ,$$

where $\alpha \in \mathbb{R}^{D \cdot |\mathcal{Y}|}$ now denotes a parameter vector defining the *softmax probabilities* and $\Phi\left(x, y \mid W\right) \in \mathbb{R}^{D \cdot |\mathcal{Y}|}$ is a sufficient statistic of $y \mid x$. If the latter is selected such that $\Phi\left(x, y \mid W\right) = \mathrm{vec}(\mathbf{e}_y \ \Psi\left(x \mid W\right)^\top)$, where $\mathbf{e}_y \in \mathbb{R}^{|\mathcal{Y}|}$ is the *one-hot* column vector having one at the position of a categorical label $y \in \mathcal{Y}$ and zero elsewhere, then the simpler form (2) is retrieved. Here, $\mathrm{vec}\left(\cdot\right)$ denotes the operator that concatenates rows of a matrix into a vector. In more complex deep learning architectures such as *sequence-to-sequence* and *attention* models [e.g., see 22], one typically designs a model-specific sufficient statistic $\Phi\left(x, y \mid W\right)$ using the decoder component. The latter takes as input the corresponding categorical label $y \in \mathcal{Y}$ along with a hidden state representation of the input sequence produced by an encoder.

A. Vicinal Risk Minimization

In empirical risk minimization, the learning algorithm selects a hypothesis that minimizes a loss/risk function with respect to the empirical distribution given by delta functions located at the training samples. In the case of acoustic models, one typically minimizes the negative log-likelihood, i.e.,

$$\mathcal{R}_{emp}\left(W, \alpha\right) = -\frac{1}{n} \sum_{i=1}^{n} \log p\left(y_i \mid x_i, \alpha, W\right) \ .$$

As the ultimate goal of a learning algorithm is to select a hypothesis that minimizes the expected risk, the vicinal risk minimization [23] aims at improved estimates by replacing the delta functions with some approximation of the density in the vicinity of training instances. More formally, the vicinal variant of the negative log-likelihood is given by

$$\mathcal{R}_{vic}\left(W, \alpha\right) = -\frac{1}{n} \sum_{i=1}^{n} \int \log p\left(y_i \mid x, \alpha, W\right) \, \mathrm{d}P_{x_i}\left(x\right) \ ,$$

where $P_{x_i}\left(\cdot\right)$ is a density estimate in the vicinity of x_i.

Data augmentation can be seen as an instance of vicinal risk minimization, where local density estimates are designed to improve the generalization properties of the empirical estimators. To illustrate this, we assume that the density estimate in the vicinity of x_i can be approximated by a mixture of Gaussians, i.e.,

$$P_{x_i}\left(x\right) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{N}\left(\mu_{ik}, \sigma_k^2 \mathbb{I}\right) \ , \qquad (3)$$

with $\mu_{ik} \coloneqq \mu_{ik}\left(x_i\right) \in \mathbb{R}^d$ and $\sigma_k \in \mathbb{R}$. In Section IV, we propose several data augmentation schemes and formally introduce these mean functions as linear transformations of training samples. The variance parameter is included here to control the signal-to-noise ratio and, thus, the level of robustness associated with each of the mixture components.

The vicinal risk induced by P_{x_i} is then given by

$$\ell_n\left(W, \alpha\right) = -\frac{1}{nK} \sum_{i,k} \mathbb{E}_{x \sim \mathcal{N}\left(\mu_{ik}, \sigma_k^2 \mathbb{I}\right)} \left[\log p\left(y_i \mid x, \alpha, W\right)\right] \ ,$$

with $1 \leq i \leq n$ and $1 \leq k \leq K$. As the integral defining the expectation operator is intractable, we resort to Monte Carlo approximation with a small number of IID samples

$$\ell_{n,m}\left(W, \alpha\right) = -\frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} \log p\left(y_i \mid x_{ij}, \alpha, W\right) \ , \qquad (4)$$

where $x_{ij} = \mu_{ik'} + \epsilon_{ik'}$ is a sample from the mixture model that corresponds to instance x_i, realized by first selecting uniformly a mixture component $k' \sim \mathcal{U}_{\{1,\dots,K\}}$ and then the offset $\epsilon_{ik'} \sim \mathcal{N}\left(0, \sigma_{k'}^2 \mathbb{I}\right)$. Due to the fact that log is a concave function, we have from the Jensen inequality that

$$\ell_n\left(W, \alpha\right) \geq -\frac{1}{n}\sum_{i=1}^{n} \log \mathbb{E}_{x \sim P_{x_i}(x)}\left[p\left(y_i \mid x, \alpha, W\right)\right] \ .$$

Hence, the vicinal risk function is an upper bound on the negative log-expected likelihoods over neighborhoods centered at training instances and minimization of such an objective is likely to promote locally smooth hypotheses.

For density estimates P_{x_i} defined with an isotropic zero-mean Gaussian distribution, Chapelle et al. [Section 3, 23] have demonstrated that the notion of vicinal risk amounts to introducing a penalty term given by the squared norm of the parameter vector, i.e., the standard ℓ_2-regularization mechanism for linear models. Thus, learning a neural network by optimizing the vicinal risk from Eq. (4) can be seen as an extension of the standard regularized risk minimization principle that works effectively for linear models.

An alternative insight into the properties incorporated into the learning algorithms via vicinal risk minimization can be obtained by introducing a temperature parameter T into the likelihood function:

$$p_T\left(y \mid x, \alpha, W\right) = \frac{\exp\left(\alpha_y^\top \Psi\left(x \mid W\right)/T\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\alpha_{y'}^\top \Psi\left(x \mid W\right)/T\right)} \ .$$

Then, we have that $\lim_{T \to 0} \ p_T\left(y \mid x, \alpha, W\right) = \xi_y$ with

$$\xi_y = \begin{cases} 1 & \text{if } y = \arg\max_{y' \in \mathcal{Y}} \ p\left(y' \mid x, \alpha, W\right) \\ 0 & \text{otherwise} \ . \end{cases}$$

From the latter expression it follows that for each mixture component from Eq. (3)

$$\begin{aligned} &\mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I})}\left[\lim_{T \to 0} \ p_T\left(y \mid x + \epsilon, \alpha, W\right)\right] \\ &= P_\epsilon\left(y = \arg\max_{y' \in \mathcal{Y}} p\left(y' \mid x + \epsilon, \alpha, W\right)\right) \end{aligned}$$

In other words, minimization of the objective in Eq. (4) can be seen as an approximation to the following problem

$$\min_{W, \alpha} \ -\sum_{i,j} \log P_{\epsilon_{ij}}\left(y_i = \arg\max_{y' \in \mathcal{Y}} p\left(y' \mid \mu_{ij} + \epsilon_{ij}, \alpha, W\right)\right),$$

which maximizes the likelihood of correct classification over neighborhoods rather than particular training instances. The latter estimator is known as *randomized smoothing* in machine learning and several recent works have provided certifiable robustness bounds for that estimator [e.g., see 24, 25].

B. Inductive Bias via Data Augmentation

In this section, we provide two further insights into the effects of data augmentation on the robustness of acoustic models. The technical derivations build on prior work by [26] that focused on data augmentation for kernel machines. We extend those observations/results to neural networks and show from another perspective that learning with augmented samples promotes vicinal smoothness of the learned models/hypotheses.

We first recall that $\alpha \in \mathbb{R}^{D \times |Y|}$ is the *softmax parameter matrix* with columns given by $\alpha_y \in \mathbb{R}^D$. The log-likelihood of label $y \in \mathcal{Y}$ conditioned on an instance $x \in \mathcal{X}$ can now be expressed using the following function

$$\begin{aligned} &\tau(\zeta\left(x\right), y) \coloneqq \alpha_y^\top \Psi\left(x \mid W\right) - \log \sum_{y' \in \mathcal{Y}} \exp\left(\alpha_{y'}^\top \Psi\left(x \mid W\right)\right) \\ &= \zeta_y\left(x\right) - \log \sum_{y' \in \mathcal{Y}} \exp\left(\zeta_{y'}\left(x\right)\right) \ , \end{aligned}$$

where $\zeta\left(x\right) \coloneqq \alpha^\top \Psi\left(x \mid W\right) \in \mathbb{R}^{|\mathcal{Y}|}$ is a vector with components $\zeta_y\left(x\right) \in \mathbb{R}$. Now, recall also that the density in the vicinity of an instance is approximated using a mixture of Gaussians (see Eq. 3), which allows for a re-parametrization via $P\left(\epsilon\right) \coloneqq P_{x_i}\left(x_i + \epsilon\right)$ for $\epsilon \in \mathbb{R}^d$.

Taking the first-order Taylor approximation of the function τ around $\zeta^{(0)}$ and passing it through the expectation operator with respect to the vicinal density $P\left(\epsilon\right)$, we obtain

$$\begin{aligned} &\mathbb{E}_\epsilon\left[\tau\Big(\zeta\left(x + \epsilon\right), y\Big)\right] = \\ &\tau\left(\zeta^{(0)}, y\right) + \nabla\tau\left(\zeta^{(0)}, y\right)^\top \mathbb{E}_\epsilon\left[\zeta\left(x + \epsilon\right) - \zeta^{(0)}\right] + \\ &o\left(\mathbb{E}_\epsilon\left[\left\|\zeta\left(x + \epsilon\right) - \zeta^{(0)}\right\|\right]\right) \ . \end{aligned}$$

Setting $\zeta^{(0)} = \mathbb{E}_\epsilon\left[\zeta\left(x + \epsilon\right)\right]$ we have $\mathbb{E}_\epsilon\left[\zeta\left(x + \epsilon\right) - \zeta^{(0)}\right] = 0$, which implies that the vicinal negative log-likelihood is approximately equal to

$$\ell_n\left(W, \alpha\right) \approx -\frac{1}{n}\sum_{i=1}^{n} \log p\Big(y_i \mid \mathbb{E}_\epsilon\left[\Psi\left(x_i + \epsilon \mid W\right)\right], \alpha\Big) \ .$$

Thus, the vicinal risk objective approximately maximizes the likelihood of the mean embedding of a training instance with respect to the density estimate P_{x_i}, which is also used as a data augmentation mechanism.

On the other hand, the second order Taylor approximation with respect to $\zeta^{(0)}$ gives

$$\begin{aligned} &\mathbb{E}_\epsilon\left[\tau\Big(\zeta\left(x + \epsilon\right), y\Big)\right] \\ &\approx \tau\left(\zeta^{(0)}, y\right) + \frac{1}{2}\mathbb{E}_\epsilon\left[\Delta_\zeta\left(\epsilon\right)^\top \nabla^2\tau\left(\zeta^{(0)}, y\right)\Delta_\zeta\left(\epsilon\right)\right] \\ &\leq \tau\left(\zeta^{(0)}, y\right) + \frac{\mu}{2}\operatorname{tr}\left(\mathbb{E}_\epsilon\left[\Delta_\zeta\left(\epsilon\right)\Delta_\zeta\left(\epsilon\right)^\top\right]\right), \end{aligned}$$

where μ is the largest eigenvalue of Hessian matrix $\nabla^2\tau\left(\zeta^{(0)}, y\right)$ and $\Delta_\zeta\left(\epsilon\right) = \zeta\left(x + \epsilon\right) - \zeta^{(0)}$. The second term is just the variance of the individual entries of the vector $\zeta\left(x + \epsilon\right) \in \mathbb{R}^{|\mathcal{Y}|}$, relative to the data augmentation distribution from Eq. (3). The latter follows from the fact that $\Delta_\zeta\left(\epsilon\right)$ is a zero-mean random variable. Now, the vicinal negative log-likelihood is approximately equal to

$$\begin{aligned} \ell_n\left(W, \alpha\right) \approx &-\frac{1}{n}\sum_{i=1}^{n} \log p\left(y_i \mid \mathbb{E}_\epsilon\left[\Psi\left(x_i + \epsilon \mid W\right)\right], \alpha\right) \\ &+ \mathbb{E}_\epsilon\left[\|\Delta_\zeta\left(\epsilon\right)\|^2_{\nabla^2\tau\left(\zeta^{(0)}, y_i\right)}\right] \leq \\ -\frac{1}{n}\sum_{i=1}^{n} \log p\left(y_i \mid \mathbb{E}_\epsilon\left[\Psi\left(x_i + \epsilon \mid W\right)\right], \alpha\right) &+ \frac{\mu}{2}\operatorname{Var}\left[\Delta_\zeta\left(\epsilon\right)\right] \ . \end{aligned}$$

From here it follows that vicinal risk minimization optimizes a lower bound on the objective consisting of the negative log-likelihood and a variance-based penalty term. The latter is defined over the space of pre-softmax vectors and relative to the random variable defined by vicinal density. Thus, the second order approximation tells us that vicinal risk minimization aims at assigning a similar conditional distribution of labels given an instance, across a neighborhood specified by the data augmentation principle. As a result, the predictions are likely to remain the same in local neighborhoods/vicinity around training points which fosters locally robust sufficient statistic.

C. A Notion of Locally Robust Sufficient Statistic

In this section, we introduce a notion of a locally robust sufficient statistic $\Psi(x \mid W)$ and provide a bound on the deviation between its values over a neighborhood in the vicinity of a training sample. In our analysis, we focus on robustness relative to a ball of constant radius centered at a training sample, which contains the high density vicinal region described by Eq. (3) in its interior. Henceforth, we will simplify our notation and denote this sufficient statistic with $\Psi(x)$ (omitting W).

A robust representation of waveform signals $\Psi(x)$ should be stable with respect to additive noise perturbations. This can be, for instance, achieved with a contraction mapping. An operator Ψ is said to be a contraction if there exists a positive constant $L < 1$ such that for all $x \in \mathcal{X}$ and $\epsilon \in \mathbb{R}^d$ with $x + \epsilon \in \mathcal{X}$

$$\|\Psi(x+\epsilon) - \Psi(x)\| \leq L \|\epsilon\| \ .$$

The contraction property is a stability notion that holds across the whole space and represents a *strong notion of robustness*. Typically, this is relaxed by requiring that the sufficient statistic is stable under small additive perturbations of the signal. More formally, for a constant $r > 0$ and all $z \in \mathcal{B}(x, r) = \{x + \epsilon \in \mathcal{X} \mid \|\epsilon\| < r\}$ it is required that $\|\Psi(z) - \Psi(x)\| < r$.

A more flexible way to express the latter notion of robustness is to assume that ϵ is a random variable that follows an isotropic Gaussian distribution, i.e., $\epsilon \sim \mathcal{N}\left(\epsilon \mid 0, \sigma^2 \mathbb{I}\right)$, and require that for all $0 < \delta < 1$ and some $r := r(\sigma, \delta) > 0$

$$\mathbb{E}_\epsilon \left[\|\Psi(x+\epsilon) - \Psi(x)\| < r\right] > 1 - \delta \ .$$

This is equivalent to requiring that for all $0 < \delta < 1$ there exists $r > 0$ such that $P_\epsilon\left(\|\Psi(x+\epsilon) - \Psi(x)\| \geq r\right) < \delta$.

Before proceeding with the bound on the deviation between values of a sufficient statistic over a neighborhood centered at a training sample, we introduce the relevant operators and notions that will be used to express this type of robustness.

Definition 1. *Let $\nabla\Psi(x) \in \mathbb{R}^{d \times D}$ be the Jacobian matrix of the sufficient statistic $\Psi(x) \in \mathbb{R}^D$ at a training sample $x \in \mathbb{R}^d$, which is given by $\nabla\Psi_{ij}(x) = \frac{\partial}{\partial x_i}\Psi_j(x)$. Let also $\nabla^2\Psi(x) \in \mathbb{R}^{d \times d \times D}$ be the Hessian tensor of the same sufficient statistic given by $\nabla^2\Psi_{ik,j}(x) = \frac{\partial}{\partial x_i}\frac{\partial}{\partial x_k}\Psi_j(x)$ with $1 \leq i, k \leq d$ and $1 \leq j \leq D$. The constants*

$$a := \operatorname{tr}\left(\nabla\Psi(x)^\top \nabla\Psi(x)\right) \quad \textit{and}$$

$$b := \sum_{j=1}^{D} \operatorname{tr}\left(\nabla^2\Psi_{**,j}(x)\right) + \operatorname{tr}\left(\nabla^2\Psi_{**,j}(x)\, \nabla^2\Psi_{**,j}(x)\right)$$

with $\Psi_{**,j}(x) \in \mathbb{R}^{d \times d}$, describe the spectra of Jacobian and Hessian tensors, thus encapsulating the geometry of the neighborhood around an instance $x \in \mathcal{X}$.

The following theorem characterizes a class of sufficient statistics that satisfy the notion of local robustness introduced above (a proof of this result is given in the appendix).

Theorem 1. *Suppose that the Hessian tensor of sufficient statistic $\Psi(x)$ is uniformly bounded for all $x \in \mathcal{X}$. Then, for all $\delta > 0$*

$$P_\epsilon\left[\|\Psi(x+\epsilon) - \Psi(x)\| < \frac{\sigma}{\delta}\left(\sqrt{a} + \sigma\sqrt{b/2}\right)\right] > 1 - \delta \ .$$

A practical meaning of this result is that the pre-softmax layer outputs that correspond to speech signals sampled from the vicinal density around a training sample concentrate in the feature space given by that layer.

IV. WaveAugment: Data Augmentation for Learning Robust Acoustic Models

In this section, we introduce four data augmentation schemes operating directly in the waveform domain with the goal of scrambling spurious correlations from the training sample and allowing for learning of robust acoustic models capable of generalizing to unseen environments. The main focus of our study is the improvement in out-of-distribution generalization when there is a mismatch between training and testing conditions that can be characterized by differences in background noise types, differences between training and run-time microphones, and room reverberation. In the remainder of the section, we provide algorithmic descriptions and motivation for the proposed data augmentation schemes based on insights from signal processing and the observed effectiveness in our empirical study (see Section VI for details). Figure 2 illustrates the effects of the proposed data augmentation schemes on the magnitude spectrum of a clean speech utterance from AURORA4 [27].

A. Band-limited White Noise

We start with an augmentation scheme that transforms an input speech waveform by adding a noise signal with support over low-frequency components only. The main motivation for this is to scramble spurious correlations that may be present in the training sample and can impede generalization to certain noise environments that affect this part of the spectrum (e.g., babble, airport, or car noise). The underlying source of randomness is white noise and we use different signal-to-noise ratios to further diversify the resulting signal corruptions.

Algorithm 1 provides a pseudo-code description of this augmentation scheme. In steps 1 and 2, we set p evenly spaced modulation frequencies ω_i and corresponding bandwidths ξ_i of Parzen filters which are defined via cosine modulations of the squared Epanechnikov window (see Section II). These are bandpass filters with support over the low-frequency part of the spectrum, with the lowest and highest frequencies specified as input to the algorithm. Following this, step 3 selects one such filter uniformly at random. The selected filter is then convolved with a white noise signal (step 5) so that the resulting additive noise vector has support over low-frequency components alone.

Algorithm 1 BAND-LIMITED WHITE NOISE

Input: audio signal $x(t)$, sampling rate f, bandpass frequency range given by $\omega_{\min}$ and $\omega_{\max}$, filterbank size p, filter support size T, SNR range given by $\gamma_{\min}$ and $\gamma_{\max}$

1: $\{\omega_i\}_{i=1}^{p} \leftarrow$ EVENLY_SPACED_MODES $(\omega_{\min}, \omega_{\max}, f, p)$
2: $\{\xi_i\}_{i=1}^{p} \leftarrow$ BANDWIDTHS $(\{\omega_i\}_{i=1}^{p}, \omega_{\min}, \omega_{\max}, f)$
3: $(\omega, \xi) \sim \mathcal{U}_{\{(\omega_1,\xi_1),\ldots,(\omega_p,\xi_p)\}}$
4: $h(t) \leftarrow$ PARZEN_FILTER (ω, ξ, T)
5: $\epsilon(t) \leftarrow (\epsilon_0 * h)(t)$ with $\epsilon_0(t) \sim \mathcal{N}(0,1)$
6: $\epsilon(t) \leftarrow$ SNR_SCALE $(\epsilon(t), \gamma)$ with $\gamma \sim \mathcal{U}(\gamma_{\min}, \gamma_{\max})$
7: **return** $x(t) + \epsilon(t)$

As the final step in the generation of band-limited white noise, the algorithm selects a signal-to-noise ratio uniformly at random from a given input range (step 6). The output of the algorithm is an additive corruption of the input speech waveform signal. In our implementation, we have used $p = 8$ Parzen filters and restricted the modulation range by setting $\omega_{\min} = 50$ and $\omega_{\max} = 800$ Hz. The corresponding bandwidths ξ_i are set to be equal and jointly cover the specified frequency band, that is, $\xi_i \approx (\omega_{\max} - \omega_{\min})/p$. For the signal-to-noise ratio range, we have opted for corruptions in the range of 8-32 dB.

We conclude with a reference to our theoretical considerations from Section III. This scheme does not involve any change to the input signal prior to adding the additive noise term. As a result, it corresponds to a Gaussian mixture component with the mean parameter given by the training sample (i.e., input signal to Algorithm 1) itself and a non-isotropic variance term (due to the convolution with the low-pass filter) that accounts for the appropriate signal-to-noise ratio (see also Eq. 3).

B. Notch Filtered Signals

In this section, we propose a data augmentation scheme based on notch filters [28] that removes certain frequencies from the input and replaces them with white noise. The main motivation behind this scheme is to bias the learning process away from certain types of spurious correlations by embedding a noise signal, which is independent of sub-phonetic labels, into frequency ranges susceptible to environment effects. There are two use cases that we would like to target with this type of augmentation. The first one deals with low-frequency interference due to microphones and the second one with scrambling of the high frequency content of the signal that is susceptible to environment effects (e.g., street or car noise).

A notch filter is defined by a frequency that sets a dip in the spectrum around which the content is eliminated. We use three-tap notch filters of the form

$$h_\omega[t] = \begin{cases} 1 & t = \pm 1\ , \\ -2\cos(\omega) & t = 0\ , \\ 0 & \text{otherwise}\ , \end{cases} \tag{5}$$

where ω is the location of the *frequency dip*. In our particular case, we would like to introduce two frequency dips, one at zero to tackle the microphone effects and another in the high frequency range (e.g., above 5 kHz).

Algorithm 2 NOISY DOUBLE-DIP COSINE NOTCH

Input: audio signal $x(t)$, sampling rate f, notch frequency range given by $\omega_{\min}$ and $\omega_{\max}$, number of high frequency notch filters p, SNR range given by $\gamma_{\min}$ and $\gamma_{\max}$

1: $\{\omega_i\}_{i=1}^{p} \leftarrow$ EVENLY_SPACED_MODES $(\omega_{\min}, \omega_{\max}, f, p)$
2: $\omega \sim \mathcal{U}_{\{\omega_1,\ldots,\omega_p\}}$ with high freq. notch filter h_ω (see Eq. 5)
3: $z(t) \leftarrow (x * h_0)(t)$ with notch filter h_0 (see Eq. 5)
4: $z(t) \leftarrow (z * h_\omega)(t)$ and $\epsilon(t) \sim \mathcal{N}(0,1)$
5: $\epsilon(t) \leftarrow$ SNR_SCALE $(\epsilon(t), \gamma)$ with $\gamma \sim \mathcal{U}(\gamma_{\min}, \gamma_{\max})$
6: **return** $z(t) + \epsilon(t)$

Algorithm 2 provides a pseudo-code description of this signal transformation. It takes as input the sampling rate along with high frequency notch range given by $\omega_{\min}$ and $\omega_{\max}$, the number of notch frequencies p, and signal-to-noise ratio range that will dictate the magnitude of white noise added to the filtered signal. In the first step, the algorithm creates a set of frequencies in the specified range from which a high frequency notch is selected uniformly at random (step 2). Following this, the notch filter with a dip at zero is convolved with the original signal (step 3). The algorithm then takes the selected high frequency dip (see step 2) and convolves the resulting (zero notched) signal with the second notch filter defined by it (step 4). As a result, the signal now has frequency dips at zero and the selected high frequency. In the next step, we generate a white noise and select a signal-to-noise ratio uniformly at random from the range provided as input to the algorithm (step 5). The output of the algorithm is an additive corruption of the notch filtered signal. As the white noise has support over the whole frequency range, this scheme injects noise signal that is independent of sub-phonetic labels at the selected *frequency dips* of the original signal and in this way fosters robustness. In our implementation, we restrict the high frequency notch range by setting $\omega_{\min} = 5,000$ and $\omega_{\max} = 8,000$ Hz. Signal-to-noise ratio range is again set to 8-32 dB.

This augmentation scheme modifies the input signal via two convolutional operators (steps 3 and 4). These are linear operators that can be realized by multiplying the input signal with a circulant matrix. In particular, there exist circulant matrices C_0 and C_ω that correspond to notch filters h_0 and h_ω (respectively) such that given an input signal x, the corresponding transformed signal is $z = C_\omega C_0 x$. In terms of Eq. (3), this means that each of the frequency dips $\{\omega_k\}_{k=1}^{p}$ defines a mixture component such that $\mu_{ik}(x_i) := \mu_{\omega_k}(x_i) = C_{\omega_k} C_0 x_i$ with $1 \leq i \leq n$ and $1 \leq k \leq p$, and where x_i denotes a training sample. The variance term in this augmentation scheme is isotropic and it accounts for the signal-to-noise ratio.

C. Wide Band-pass Filtered Signals

Spurious correlations due to microphone effects are challenging to address when learning directly in the waveform domain. The main goal in this section is to devise an augmentation scheme that could be effective in dealing with adverse conditions related to mismatch between training and test microphones. Such effects can be characterized by suppression of content in high and low frequency regions of the spectrum.

Algorithm 3 NOISY WIDEPASS

Input: audio signal $x(t)$, sampling rate f, bandpass frequency range given by $\omega_{\min}$ and $\omega_{\max}$, filterbank size p, filter support size T, SNR range given by $\gamma_{\min}$ and $\gamma_{\max}$

1: $\{\omega_i\}_{i=1}^{p} \leftarrow$ EVENLY_SPACED_MODES $(\omega_{\min}, \omega_{\max}, f, p)$
2: $\{\xi_i\}_{i=1}^{p} \leftarrow$ WIDE_BANDWIDTHS $(\{\omega_i\}_{i=1}^{p}, \omega_{\min}, \omega_{\max}, f)$
3: $(\omega, \xi) \sim \mathcal{U}_{\{(\omega_1,\xi_1),\ldots,(\omega_p,\xi_p)\}}$
4: $h(t) \leftarrow$ PARZEN_FILTER (ω, ξ, T)
5: $z(t) \leftarrow (x * h)(t)$ and $\epsilon(t) \sim \mathcal{N}(0,1)$
6: $\epsilon(t) \leftarrow$ SNR_SCALE $(\epsilon(t), \gamma)$ with $\gamma \sim \mathcal{U}(\gamma_{\min}, \gamma_{\max})$
7: **return** $z(t) + \epsilon(t)$

Algorithm 4 NOISY RIR

Input: audio signal $x(t)$, sampling rate f, set of 3D room configurations $\mathcal{C}$, set of wall materials $\mathcal{M}$, source to microphone distances given by $d_{\min}$ and $d_{\max}$, SNR range given by $\gamma_{\min}$ and $\gamma_{\max}$

1: $c \sim \mathcal{U}_{\{c_1,\ldots,c_{|\mathcal{C}|}\}}$ with $\mathcal{C} = \{c_i\}_{i=1}^{|\mathcal{C}|}$
2: $m \sim \mathcal{U}_{\{m_1,\ldots,m_{|\mathcal{M}|}\}}$ with $\mathcal{M} = \{m_i\}_{i=1}^{|\mathcal{M}|}$
3: mic $\leftarrow$ vec (u_1, u_2, u_3) with $u_i \sim \mathcal{U}(0, c[i])$
4: $h \leftarrow$ RIR (c, m, mic, d) with $d \sim \mathcal{U}(d_{\min}, d_{\max})$
5: $z(t) \leftarrow (x * h)(t)$ and $\epsilon(t) \sim \mathcal{N}(0,1)$
6: $\epsilon(t) \leftarrow$ SNR_SCALE $(\epsilon(t), \gamma)$ with $\gamma \sim \mathcal{U}(\gamma_{\min}, \gamma_{\max})$
7: **return** $z(t) + \epsilon(t)$

Hence, we would like to pass the input signal through a wide band-pass filter that removes information at low and high frequencies, and replace that information by white noise, which is independent of labels, and thus biases the learning process away from over-fitting the microphone effects.

Algorithm 3 provides a pseudo-code description of this signal transformation. In the first step, we create a set of p evenly spaced modulation frequencies covering the frequency range provided as input. Then, the algorithm creates a set of p bandwidths (step 2) such that the resulting filters are with wide band-pass properties (see step 4). Following this, we select one of these filters uniformly at random (steps 3 and 4) and convolve it with the input signal (step 5), thus retaining only the information covered by the filter support. In the final step, a white noise vector is scaled such that the additive corruption of band-passed signal has the appropriate signal-to-nose ratio, sampled uniformly from the range given as input to the algorithm (step 6). We simulate this algorithm with $p = 8$ bandpass Parzen filters (see Section II) with the support over frequency range 50-7950 Hz. The bandwidths are selected to follow the Mel-scale, that is, the bandwidth ξ_i of the filter with centre frequency ω_i is set to be approximately equal to the width of the band at that same frequency on the Mel-scale. In the case of relatively small number of filters, that ensures that filters have wider bandwidths. For the signal-to-noise ratio we use the range of 8-32 dB.

This augmentation scheme alters the input signal via convolution with a band-pass filter. The operation can be realized by multiplication with a circulant matrix, i.e., for a band-pass filter h there exists a circulant matrix C_h such that $z = C_h x$. In terms of Eq. (3), each of the band-pass filters $\{(\omega_k, \xi_k)\}_{k=1}^{p}$ from steps 1-4 in Algorithm 3 defines a mixture component such that $\mu_{ik}(x_i) \coloneqq \mu_{\omega_k,\xi_k}(x_i) = C_{\omega_k,\xi_k} x_i$ with $1 \leq i \leq n$ and $1 \leq k \leq p$, and where x_i denotes a training sample. As in the previous scheme, the variance term is isotropic and accounts for the signal-to-noise ratio.

D. Reverberation Effects

This section covers a data augmentation scheme based on room impulse response modeling. Reverberations introduce spurious correlations in speech signals as the result of a large distance between speakers and a microphone [13]. In particular, speech signal reaches a microphone via a direct line-of-sight path, followed by first, second and higher order reflections off the walls and other objects. These reflections, referred to as reverberation, can be represented as linear convolutions of the speech signal with the room impulse response. There are two stages characteristic to this process, early reflections that typically occur within 50 ms, followed by a dense series of higher-order reflections called late reverberations [13, 29]. The challenging aspect of this process is the fact that late reverberation is non-stationary and is, thus, not addressable by standard noise compensation techniques such as vector Taylor series [13]. The aim of the augmentation scheme introduced in this section is to scramble spurious correlations that can occur as a result or distant-talking and reverberations. *Thus, we are not interested in designing noise compensation mechanisms but are willing to delegate this task to the network itself with inductive bias guided by examples of linear convolutions of input signals with different room impulse responses.*

Algorithm 4 provides a pseudo-code description of this augmentation scheme. First, we select (uniformly at random) one of the possible room configurations (step 1) and a wall material setting (step 2). Following this, the algorithm samples a location of the microphone based on the selected room dimensions (step 3) along with the distance to an acoustic source (step 4). The selected parameters are then passed to the *pyroomacoustics* library [30], which generates a room impulse response. The algorithm convolves the input speech signal with the selected room impulse response and further corrupts the resulting signal with an additive white noise, appropriately scaled to respect the selected signal-to-noise ratio (steps 5-6). The reason for adding white noise to the reverberated signal is to avoid having repeated noise types in the training samples as deep learning models may learn to remove them, defeating the purpose of scrambling spurious correlations introduced by distant-talking. We note here that white noise has been previously combined with reverberated signals in [31].

In our implementation of this data augmentation scheme, we use the following three room configurations (in meters)

$$\mathcal{C} = \{[4, 4, 2.5], [10, 10, 3.5], [2.5, 1.5, 1.5]\}$$

along with five types of wall materials $\mathcal{W} = \{$*hard surface, marble floor, wooden door, glass window, hairy carpet*$\}$. For each of the material types, we select one of the four possible scattering modes [30]: *none, rpg-skyline, classroom tables, and*

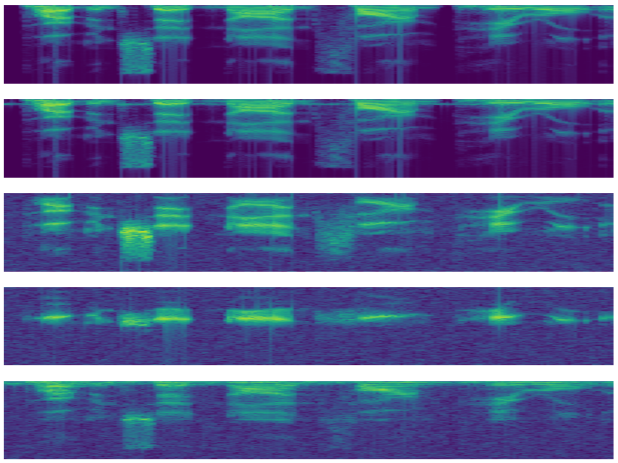

Fig. 2: The figure illustrates perturbations to the magnitude spectrum of an utterance generated by the proposed algorithms. The panels (ordered from top to bottom) depict a clean speech utterance along with its perturbations by Algorithm 1-4, respectively.

rectangular prism boxes. The distance to the microphone is selected uniformly at random in the range between 3 cm and 3 m. As in previous cases, white noise is added with the resulting signal-to-noise ratio in the range 8-32 dB.

Similar to the two previously covered augmentation schemes, the input signal is altered using the convolution operator defined by a RIR filter. In the context of Eq. (3) and our theoretical analysis this means that each RIR filter corresponds to a mixture component. More formally, for a RIR filter h there exists a circulant matrix C_h such that $z = C_h x$. Thus, mixture component means (see Eq. 3) that arise as a result of using this augmentation scheme are given by $\mu_{ik}(x_i) \coloneqq \mu_{h_k}(x_i) = C_{h_k} x_i$ with $1 \leq i \leq n$ and a set of RIR filters $\{h_k\}_k$.

V. Related Work

Data augmentation has been widely used in the past to improve performance of acoustic models on different tasks. The process of augmenting the training sample is typically done in an improvised manner using well established heuristics (e.g., masking blocks of frequency channels, speed and tempo perturbations, additive noise, etc.) and little understanding of the underlying machine learning principles. The latter aspect is a contribution of our work compared to previously devised augmentation schemes for improving robustness to distribution shifts and out-of-distribution generalization performance (e.g., see Section III). Previous data augmentation approaches can be divided into three classes: *i*) acoustic data perturbations, *ii*) text-to-speech augmentation, and *iii*) augmentation with unsupervised data (semi-supervised and self-supervised learning).

The first class of approaches aims at reducing the divergence between training and test conditions/distributions by transformations of the inputs that tackle common sources of mismatch such as additive and channel noise, reverberations, microphone types, speaker variations, etc. The perturbations are typically label-preserving and do not require additional alignments (e.g., when learning with hybrid HMM-DNN models). This process of adding label-preserving noisy data with different types and levels of noise to the original training samples is known as *multi-condition/style training* [32] and it is used frequently for learning of robust acoustic models [e.g., see 1, 14, 27]. When the noisy data is generated using a database with a limited number of naturally occurring environmental noise signals the approach may fail to generalize beyond the training distribution. More specifically, it can happen that a neural network *memorizes* all the noise signals and learns to *subtract* them from the augmented data [e.g., see 33]. Indeed, the noise in multi-condition training is typically constant — the utterances are corrupted prior to training and frames are repeatedly (in each epoch) presented to a learning algorithm with identical corruptions. Hence, it can be challenging for the network to differentiate between signal and spurious correlation because they might be coupled throughout the training process. It is for this reason that our augmentation schemes (presented in Section IV) rely on additive Gaussian noise with various signal-to-noise ratios to generate diverse label-preserving input perturbations. Moreover, our augmentation approaches are amenable to *online* data perturbations where at the onset of an epoch one can randomly select an additive corruption or proceed with the original utterance (see Sections VI-D and VI-E). SpecAugment [3] is a recently proposed highly effective data augmentation scheme from this class. It operates on the log-mel-spectrogram of the input audio and consists of steps such as feature warping, masking blocks of frequency channels, and masking blocks of time steps. A shortcoming is the restriction to filterbank features, which precludes its use in other types of acoustic models (e.g., truly end-to-end speech recognition systems). In contrast to this, the augmentation schemes proposed in Section IV apply to all types of acoustic models and to some extent model the policies characteristic to [3]. In particular, Algorithm 3 transforms the input signal and replaces blocks of frequency channels with white noise which is independent of labels and can, thus, scramble spurious correlations in that part of the spectrum. Related to our augmentation schemes (i.e., Algorithm 1) is the bandpass noise generation approach proposed in [4]. More specifically, that approach takes a database of naturally occurring sounds along with a bandpass filterbank to generate a set of bandlimited noise signals. The empirical results in [4] report a 7% relative improvement in generalization to unseen noise conditions (far-field device and car noise). In contrast to this approach, Algorithm 1 relies exclusively on independent Gaussian samples that scramble spurious correlations in the low-frequency region. Another set of transformations directly relevant for our work are room impulse response (RIR) perturbations. Previous work reports improvement in distant and noisy speech as a result of this augmentation scheme [e.g., see 13, 29]. We extend this line of work by combining additive Gaussian noise with RIR perturbation, thus tackling the potential memorization effects when learning with an over-parameterized neural network (see Algorithm 4). Speed and tempo perturbations are also frequently used transformations of the input signal that can improve generalization in conversational speech [e.g., see 34]. A potential issue with this approach is the change in the duration of an utterance, requiring re-alignment of target labels (e.g., senones in hybrid HMM-DNN models) for a transformed audio that might not be supported by the training distribution (thus resulting in poor/noisy training labels). Here

it is important to note that speed perturbations were designed to emulate the vocal tract length perturbations (VTLP) that aim at improving robustness by adding synthetic speaker variations [e.g., see 34, 35]. Interestingly, it has been observed in [35] that the most effective perturbations are those applied to linear spectrograms, which is in line with the theoretical principles outlined in [Section II.C, 36] and realized by the augmentation schemes proposed in Section IV. Namely, in all of the cases the transformations of the inputs aim at removing spurious correlations present in linear spectrograms.

The second class of approaches aims at improving robustness of acoustic models by synthetically adding data using speech synthesis. This allows for training of models with utterances that were not part of the original training sample. Thus, the ultimate goal of these approaches is not to achieve robustness to novel acoustic conditions but to provide further flexibility when it comes to language constructions. Prominent augmentation schemes from this class are the approaches proposed by [37] and [38]. A common issue with this class of approaches is the lack of high quality training data (i.e., audio recordings), which manifests itself especially in low-resource languages. Another issue is that for many low-resource languages it may be difficult to find textual data in the amount typically used for building language models (e.g., for English or Mandarin).

The third class of approaches aims at improving robustness by leveraging large amounts of unlabeled data using semi-supervised and self-supervised learning. For example, in [39] an approach for acoustic models operating with low-resource languages has been devised that combines semi-supervised learning and vocal tract length perturbations. Perhaps the most prominent approach from this class of learning algorithms is the group of neural architectures known as WAV2VEC [40, 41], which rely on self-supervised learning and large amounts of unlabeled data to extract robust representations of sub-phonetic units. Typically, the learned representation are adapted to the task at hand with additional training using a small amount of labeled data. The main advantage of this class of approaches is in the fact that unlabelled data is typically easy to obtain, e.g., news broadcasts covering various different speaker and noise conditions [39]. A potential shortcoming of *unsupervised* approaches can be the lack of high quality transcriptions, which are important for transferring the learned representations to the task at hand. In [39] it has been observed that the lack of transcriptions can limit gains from some semi- and unsupervised approaches such as those based on discriminative training [42] and speaker adaptations based on discriminative criteria [43]. Recent empirical results, however, show that self-supervised learning can be effective in low-resource ASR [40, 41].

VI. EXPERIMENTS

The ultimate goal of our empirical evaluation is to assess the effectiveness of the proposed augmentation techniques (Section IV) and underlying theoretical principles (Section III) in a controlled setting with a significant mismatch between training and test conditions. To this end, we perform a series of experiments on AURORA4 [27] and AMI [44] datasets with the help of the corresponding Kaldi recipes [45].

AURORA4 is a well known benchmark dataset that has been assembled for the purpose of evaluating the effectiveness of acoustic models in out-of-distribution generalization. In particular, the benchmark includes a Kaldi recipe/setting known as *clean-condition training* with a set of clean speech training utterances recorded using a high-quality close-talking *Sennheiser* microphone and a number of test utterances recorded using one of 18 different microphones. While some of the test utterances are recorded in clean conditions, the majority of them are noisy and cover various different noise conditions such as street traffic, train station, car, babble, restaurant, and airport. The signal-to-noise ratio (SNR) in these test recordings ranges from 5 to 15 dB. Hence, the clean-condition Kaldi recipe for AURORA4 allows for assessing two modes of out-of-distribution generalization: *i*) robustness to different types and levels of noise unseen during training, and *ii*) robustness relative to bias introduced by microphones and recording devices. In addition to this, the benchmark also includes a Kaldi recipe/setting known as *multi-condition training* that has been designed such that the conditions (i.e., noise types and microphones) present in training match those for the test utterances. More specifically, the training sample in this setting includes recordings from Sennheiser and other microphones, and it covers various different noise types [27]. An important difference between acoustic conditions present in the two samples for this setting is in the range of signal-to-noise ratios (10-20 dB in training vs 5-15 dB in test utterances). Here, it is important to note that throughout our empirical evaluation we maintain the mismatch in the signal-to-noise ratios between training data (detailed below) and the test conditions of AURORA4.

In our augmentation schemes, we are not relying on any of the naturally occurring noise types such as the ones present in test samples of AURORA4. Thus, a statistically significant improvement over clean-condition training, measured via the word error rate (WER) on unseen noisy tests samples, would imply that the proposed approach generalizes to unseen acoustic conditions. The alternative multi-condition training recipe allows us to learn a competitive baseline model with an identical architecture and can, thus, be used to assess the potential of the proposed augmentation schemes and underlying principles for learning noise-robust acoustic models.

The second benchmark dataset AMI comes with different Kaldi recipes for data recorded using headset and distant microphones. Again, we illustrate the utility of the proposed algorithms in out-of-distribution generalization, this time generalizing from headset recorded to distant-talking speech.

In all of our experiments, we train a context dependent model based on frame labels (i.e., HMM state ids) generated using a triphone model from Kaldi [45] with 25 ms frames and 10 ms stride. The data splits (training/development/evaluation) are identical to the ones from the corresponding Kaldi recipes. In the preprocessing step, we assign the Kaldi frame label to a 200 ms long segment of raw speech centered at the original Kaldi frame. The Parzen convolution block is initialized by taking the modulation frequencies to be equidistant in mel-scale. The bands of filters are initialized as in FBANK features. For convolutional and dense blocks in our network, we employ the Xavier initialization scheme [46] with magnitude 0.005. While

TABLE I: The table reports the word error rates obtained on AURORA4 using the Kaldi clean-condition recipe. To assess the effectiveness of the proposed augmentation schemes, the training fold has been replicated by Algorithms 1-4. The relative performance degradation/improvement is computed with respect to the second column (colored in orange), which includes all replicas of the training fold. The symbol – implies that the corresponding replica of the clean-condition training fold has been excluded from the training set in that simulation. In the rightmost column we report the results using the training set from the Kaldi clean-condition recipe alone. The column labelled DETR. reports the relative deterioration in the word error rate as a result of removing the contribution of an augmentation scheme from the training recipe — the entries colored in red reflect the negative effects on the performance as a result of training without that particular vicinal density characterization.

TEST SAMPLE	CLEAN BANDLIM. NOTCH WIDEPASS RIR GAUSS	CLEAN - NOTCH WIDEPASS RIR GAUSS		CLEAN BANDLIM. - WIDEPASS RIR GAUSS		CLEAN BANDLIM. NOTCH - RIR GAUSS		CLEAN BANDLIM. NOTCH WIDEPASS - GAUSS		CLEAN BANDLIM. NOTCH WIDEPASS RIR -		CLEAN - - - - -	
	ERR.	ERR.	DETR.	ERR.	DETR.	ERR.	DETR.	ERR.	DETR.	ERR.	DETR.	ERR.	DETR.
A. SUMMARY OVER CLEAN SPEECH WITH TRAINING MICROPHONES													
CLEAN	2.58	2.58	–	2.56	1%	2.58	–	2.47	4%	2.54	2%	**1.96**	24%
B. SUMMARY OVER NOISY SPEECH WITH TRAINING MICROPHONES													
CAR	3.53	3.61	2%	4.04	14%	3.92	11%	3.89	10%	**3.36**	5%	15.92	351%
BABBLE	6.58	9.86	50%	5.98	9%	**5.66**	14%	6.58	–	5.74	13%	16.66	153%
RESTAURANT	7.64	8.16	7%	7.64	–	**7.38**	3%	8.41	10%	7.45	2%	15.99	109%
STREET	**7.04**	7.23	3%	8.43	20%	7.83	11%	8.14	16%	7.32	4%	25.67	265%
AIRPORT	**6.26**	9.21	47%	6.41	2%	6.44	3%	6.46	3%	6.63	6%	12.67	102%
TRAIN	7.06	**6.65**	6%	8.16	16%	7.73	9%	8.89	26%	7.29	3%	26.56	276%
B. AVERAGE	6.35	7.45	17%	6.78	7%	6.49	2%	7.06	11%	**6.30**	1%	18.91	198%
C. SUMMARY OVER CLEAN SPEECH WITH DIFFERENT MICROPHONES (UNSEEN DURING TRAINING)													
CLEAN	7.79	7.58	3%	8.44	8%	12.29	58%	**7.49**	4%	8.03	3%	19.47	150%
D. SUMMARY OVER NOISY SPEECH WITH DIFFERENT MICROPHONES (UNSEEN DURING TRAINING)													
CAR	10.46	**7.85**	25%	12.46	19%	15.32	46%	9.98	5%	8.78	16%	34.32	228%
BABBLE	16.22	17.60	9%	16.59	2%	19.04	17%	18.70	15%	**15.37**	5%	38.93	140%
RESTAURANT	18.16	**18.08**	–	19.63	8%	21.65	19%	20.79	14%	18.79	3%	36.52	101%
STREET	18.74	**16.63**	11%	20.74	11%	22.73	21%	20.51	9%	17.21	8%	49.32	163%
AIRPORT	**16.50**	17.37	5%	16.98	3%	20.31	23%	18.57	13%	16.66	1%	35.59	116%
TRAIN	19.15	**16.44**	14%	20.94	9%	22.92	20%	21.32	11%	18.48	3%	48.07	151%
D. AVERAGE	16.54	**15.66**	5%	17.89	8%	20.33	23%	18.31	11%	15.88	4%	40.46	145%
SUMMARY OVER ALL 14 TEST SAMPLES													
AVERAGE	10.55	10.63	1%	11.36	8%	12.56	19%	11.59	10%	**10.26**	3%	26.98	156%

the convolutional blocks are initialized with the factor type *in*, the dense blocks use the *avg* type. The feature extraction layers (i.e., Parzen and convolutional parameters) are updated using the RMSPROP optimizer with the initial learning rate set to 0.0008. The multi-layer perceptron blocks are updated using stochastic gradient descent with the initial learning rate set to 0.08. A similar combination of optimizers (all network parameters are optimized jointly) was used in [5, 7]. When training using data augmentation, we decrease the learning rates by a factor of 2 at the end of an epoch, apart from the two initial epochs, and repeat this until the completion of training. We use minibatches of 512 samples and terminate the training process after 8 epochs. When training in the original clean- and multi-condition settings, we terminate the training process after 25 epochs and decrease the learning rates by a factor of 2 at the end of an epoch if the relative improvement in the classification accuracy on the validation fold is below 0.1%.

A. Impact of the Augmentation Schemes on Robustness

In the first set of experiments, the goal is to estimate the impact of individual augmentation schemes (Algorithms 1- 4) on the robustness to unseen noise conditions. To this end, we generate a transformed set of training utterances from the Kaldi clean-condition recipe using each of the augmentations schemes, along with a set corrupted using the Gaussian noise alone. The reason for including the latter corruption is to establish whether there are spurious correlations that are complementary to the proposed augmentation schemes and addressable via plain additive Gaussian noise. We refer to the utterances produced by Algorithms 1- 4 (respectively) as BANDLIMITED, NOTCH, WIDEPASS, and RIR. In the first experiment, we train the PARZNETS 2D model [7] using all the available data (CLEAN, Algorithms 1- 4, and additive GAUSS perturbations). The second column in Table I summarizes the performance of the model across different test sets (A: clean speech with training microphones, B: noisy speech with training microphones, C: clean speech with different microphones, and D: noisy speech with different microphones). In the remainder of this subsection, we will use the word error rate obtained in this way (i.e., WER 10.55%) as a reference and quantify the influence of an augmentation scheme relative to it by training the identical model without the corresponding training sample contribution.

Algorithm 1 (BANDLIM.)*:* The third column in Table I summarizes the results obtained by removing the contribution of this augmentation scheme from the training sample. The sub-column labeled as ERR. reports the word error rate that has slightly increased (1% relative) as a result of this intervention. The augmentation scheme contributes to significant improvement (colored in red) in word error rate on noisy samples recorded using the training microphones (babble and airport noise, 50% and 47% relative). This comes at the expense of picking up spurious correlations that can be associated with microphones, resulting in performance degradation on sample D (car, train, and street noise are affected the most). Overall, the augmentation scheme can help with spurious correlations specific to some types of additive noise.

TABLE II: The word error rates (%) obtained on different test sets of AURORA4 with various training settings.

TRAINING SETTING \ INPUT TYPE	RAW SPEECH						STANDARD FEATURES			
KALDI CLEAN-CONDITION RECIPE	✓	✓			✓					
KALDI MULTI-CONDITION RECIPE			✓	✓		✓	✓	✓	✓	✓
AUGMENTATION (ALGORITHMS 1- 4)		✓								
TEST SET \ NEURAL ARCHITECTURE	PARZNETS 2D			SINCNET	CLDNN [47]		MFCC MLP	FMLLR MLP	VDCNN [48]	OCTCNN [49]
A: CLEAN SPEECH & TRAIN MIC.	1.96	2.54	2.32	3.12	3.17	3.19	4.28	3.34	3.27	2.32
B: NOISY SPEECH & TRAIN MIC.	18.91	6.30	4.38	5.97	33.34	6.08	7.44	6.27	5.61	4.73
C: CLEAN SPEECH & DIFFERENT MIC.	19.47	8.03	4.30	5.68	16.16	6.57	8.73	5.74	5.32	4.24
D: NOISY SPEECH & DIFFERENT MIC.	40.46	15.88	12.73	16.58	45.67	14.06	18.71	16.04	13.52	13.57
AVERAGE WER	26.98	10.26	7.80	10.29	35.24	9.33	12.14	10.21	8.81	8.31

*Algorithm 2 (*NOTCH*):* The fourth column in Table I quantifies the performance for this setting, with the notch contribution removed from the training sample. Overall, there is an increase in the word error rate which signifies the importance of this augmentation scheme (8% relative over the whole test fold). The augmentation scheme can help with spurious correlations introduced by all the noise types considered, except for babble noise. Moreover, the improvement in robustness is consistent and independent of the microphones used.

*Algorithm 3 (*WIDEPASS*):* The fifth column in Table I summarizes the results with the WIDEPASS contribution removed from the training sample. The empirical results indicate that this is the most effective augmentation scheme for dealing with spurious correlations (removal results in increased word error rate, 19% relative). When it comes to dealing with microphone effects, this is the most effective strategy among the ones considered, with 58% and 23% relative improvement on clean and noisy speech, respectively. In comparison to other sampling schemes, WIDEPASS augmentation is extremely effective in dealing with spurious correlations attributed to noisy speech with different microphones (see the results for sample D).

*Algorithm 4 (*RIR*):* The sixth column in Table I summarizes the results when the RIR contribution is removed from the training sample. Overall, this augmentation scheme contributes to an approximately 10% relative improvement in the WER. It is quite effective on noisy samples, independently of the microphone used. However, this comes at the expense of a slight performance degradation on clean speech that is recorded using different microphones (4% relative). This means that the augmentation scheme can scramble useful associations between speech frames and sub-phonetic units, which can be remedied via other schemes.

GAUSS*:* The goal of this experiment is to quantify the influence of additive Gaussian noise when coupled with more structured transformations of the inputs (i.e., Algorithms 1-4). The empirical results in column seven of Table I show improvement in performance when the contribution of this transformation is *removed* from the training sample. This means that further corruption with unstructured additive Gaussian noise does not help with spurious correlations. On the contrary, it degrades the performance by removing useful associations between inputs and corresponding labels.

We conclude by comparing the best model trained with the help of data augmentation (column seven in Table I) to the one trained using the Kaldi clean-condition recipe alone (see the last column in Table I). Our empirical evidence indicates a *significant performance improvement* ($> 150\%$ relative), with augmented model achieving WER 10.26% vs. 26.98% obtained using the clean-condition training.

B. Potential of the Augmentation Schemes Relative to Matching of Acoustic Conditions between Training and Test Folds

In this section, we evaluate the potential of the proposed augmentation schemes for out-of-distribution generalization and robustness to unseen noise conditions. To this end, we exploit the aforementioned multi-condition recipe/setting for AURORA4 and compare the numbers obtained in (augmented) clean-condition training to the ones reported in [7] for the former setting. Table II summarizes the results of this comparison.

The results indicate that when training with the clean-condition recipe, augmented using schemes from Section IV, one can learn acoustic models that are competitive with those learned using the multi-condition training recipe, for which an *explicit matching* (with respect to the noise types and used microphones) between train and test conditions has been performed. Here it is important to point out that WER achieved by PARZNETS 2D in the multi-condition setting is difficult to beat by purely acoustic models trained via the clean-condition recipe (i.e., that is one of the most competitive baselines), given the *explicit matching* of training and test conditions.

Our discussion concludes with the observation that the performance of the augmented PARZNETS 2D model is also competitive with recently proposed highly effective feedforward architectures based on standard non-adaptive features, trained using a *matched* sample provided by the multi-condition recipe. We leave it for future work to further tune the proposed augmentation schemes, which have been simulated with modest number of bandpass filters and hyper-parameters.

C. Robustness Relative to Perturbations of the Test Folds

This set of experiments is motivated by the considerations in Section III, where we demonstrated that perturbations of an input speech frame sampled from the vicinal density will concentrate in the feature space given by the pre-softmax layer of a neural network. Moreover, the embeddings of perturbations will concentrate in that space around the vector representing the input frame. As a result, a neural network will assign similar conditional distributions of labels given a speech frame for a neighborhood specified by the data augmentation principles. This, in particular, refers to the concentration bound from Theorem 1 and our derivations for inductive bias (see Section III-B). Thus, if we perturb a test speech frame with the proposed augmentation schemes then the corresponding

TABLE III: The stability assessment for the PARZNETS 2D models trained using: *i*) the Kaldi clean-condition recipe and proposed data augmentation schemes, and *ii*) the Kaldi multi-condition recipe. The evaluation is performed using the original test fold of AURORA4 and two perturbation settings realized by data augmentation. The results are summarized across groups of test sets A-D, as described in Table II.

	A	B	C	D	AVG	↓ REL
PARZNETS 2D						
KALDI CLEAN-CONDITION RECIPE & ALGORITHMS 1-4						
ORIGINAL	2.54	6.30	8.03	15.88	10.26	–
SMOOTHED (X 4)	2.80	6.86	8.05	17.63	11.27	9.8%
SMOOTHED (X 40)	2.91	7.18	8.14	18.14	11.64	13.5%
PARZNETS 2D						
KALDI MULTI-CONDITION RECIPE						
ORIGINAL	2.32	4.38	4.30	12.73	7.80	–
SMOOTHED (X 4)	2.62	6.15	4.86	17.17	10.53	35.0%
SMOOTHED (X 40)	2.71	6.48	5.10	17.66	10.90	39.7%

conditional probability vectors should be aligned in robust models. To assess this, we take the best performing model that employs the proposed augmentation schemes (column seven in Table I) and evaluate its stability relative to perturbations of the test utterances. We also do the same experiment for the model obtained via multi-condition training. When comparing the performance of the two models, we will use the relative degradation in word error rates as a result of hypothesis smoothing over neighborhoods assigned to test utterances.

When assessing the stability, we generate s perturbations of a test speech frame $x \in \mathcal{X}$ and proceed to decoding with the average conditional probability over perturbations, i.e.,

$$q\left(y \mid x\right) \coloneqq \frac{1}{s} \sum_{i=1}^{s} p\left(y \mid \mathcal{T}_i\left(x\right), \alpha, W\right) \ ,$$

where $\mathcal{T}$ denotes a transformation from Algorithms 1- 4. We note here that the training fold in the Kaldi clean speech recipe for Aurora4 consists of clean speech/utterances alone. The test fold consists of noisy speech (unseen during training), as described in the first column of Table 1. In this experiment, we are interested in what happens around test points and whether the conditional probabilities of labels given frames are aligned in the vicinity of test utterances. As there is already a strong discrepancy between training and test folds, the fact that we apply Algorithms 1-4 to obtain realistic samples from the neighborhoods of tests utterances does not help our approach in any way. As evidenced with the model trained on clean speech alone (see Table I, WER 26.98%), it is challenging not to fail on the original test utterances let alone their perturbations.

Table III summarizes the results of these experiments. The word error rates of the two models over original test folds (without test perturbations) are reported in rows labeled as ORIGINAL. In rows labeled as SMOOTHED we report the error rates for two sets of experiments: *i*) adding a single perturbation for each of the four augmentation algorithms (denoted with $\times 4$), and *ii*) adding 10 perturbations for each of the four augmentation algorithms (denoted with $\times 40$).

Our empirical results demonstrate that the acoustic model learned by training using the proposed augmentation schemes exhibits a fair amount of robustness, with only a 13.5% relative performance degradation under significant perturbations (see SMOOTHED $\times 40$ under CLEAN-CONDITION in Table III) of the already challenging test fold (due to the mismatch between training and test conditions). We note that this is only a minor degradation relative to the one between models obtained via clean- and multi-condition Kaldi recipes (see Tables I and II).

We have repeated the same experiment for the model obtained using the multi-condition recipe/setting (WER 7.80%) and observed that there is a 39.7% relative performance degradation (see SMOOTHED $\times 40$ in Table III), which indicates that the model learned using our augmentation schemes might be a better choice for unseen noise environments.

D. Robustness Relative to FBANK and SPECAUGMENT

In this section, we evaluate the effectiveness of the proposed augmentation schemes relative to the SPECAUGMENT baseline [3]. As our focus is on robustness relative to different types and levels of additive noise unseen during training as well as microphone effects, we again run experiments on AURORA4 using the Kaldi clean-condition recipe. SPECAUGMENT is designed to operate on log-mel-spectrograms and in order to have an objective assessment of its effectiveness we simulate the experiment using FBANK features. We extract in total 64 features per frame using 25 ms long frames and 10 ms stride between them. This is the feature extraction setting identical to the one used by HMM-DNN model supplying the alignments. To simplify the experimental setting, we opt for a multi-layer perceptron (MLP) with five hidden layers and RELU activation function as our DNN model. After linear operators in each of the hidden layers we apply batch normalization and Bernoulli dropout with $p = 0.15$.

In a typical training regime with schemes such as SPECAUGMENT, the training utterances are corrupted at the onset of each epoch. We follow this procedure and with uniform probability ($p = 0.2$) decide whether to retain the original utterance or perturb it via Algorithms 1-4. For SPECAUGMENT we apply a similar rule and retain the original utterance approximately 20% of the times. We have opted for context size of 5 frames around a center frame and, thus, the input to MLP consists of 11 successive frames. As a result of this, we simulate SPECAUGMENT with the maximal time mask size of 10 frames and select at most 5 such masks per utterance. For frequency masking we have experimented with several settings and selected the maximal frequency mask size of 16 channels. We apply one such mask per utterance. The setting described here for SPECAUGMENT is in line with its adaptation to feedforward models [e.g., see 50].

There are several possible confounding factors when assessing the effectiveness of augmentation schemes in this setting. For example, FBANK features compress information from waveform signals and this can negatively impact robustness. In addition to this, there are strategies that can mitigate some effects of additive perturbations such as utterance level mean normalization and signal pre-emphasis — these can preclude the actual effectiveness of a data augmentation strategy. To account for the latter two factors, we perform experiments with and without them. A detailed description of the experiments is provided in Table IV. Our empirical results indicate a clear improvement as a result of employing the proposed augmentation schemes. More specifically, in

TABLE IV: The word error rates (%) obtained on different test sets of AURORA4 by training a MLP model using Kaldi clean-condition recipe with filterbank features. The columns labelled CLEAN report results using the data in the original Kaldi recipe. Other columns report the results of experiments where at the onset of each epoch with probability $p = 0.2$ the original utterance is retained, otherwise it is perturbed using the listed augmentations schemes.

	KALDI CLEAN-CONDITION RECIPE			+ALGORITHMS 1-4			+SPECAUGMENT (MAG. SPECT.)			+SPECAUGMENT (FBANK)		
PRE-EMPHASIS			✓			✓			✓			✓
UTTERANCE NORMALIZATION		✓	✓		✓	✓		✓	✓		✓	✓
A: CLEAN SPEECH & TRAIN MIC.	3.68	3.42	3.27	4.80	4.17	3.92	5.27	3.49	3.72	5.06	4.18	4.04
B: NOISY SPEECH & TRAIN MIC.	28.71	14.78	12.02	16.91	9.67	9.31	31.50	14.62	13.49	37.53	16.10	15.76
C: CLEAN SPEECH & DIFFERENT MIC.	37.23	17.00	18.20	21.89	9.36	9.77	35.29	17.07	17.26	36.60	14.57	15.45
D: NOISY SPEECH & DIFFERENT MIC.	54.05	31.54	29.01	36.58	20.28	19.72	55.22	29.90	28.96	58.92	30.51	30.91
AVERAGE WER	38.39	21.39	19.12	24.83	13.80	13.42	40.06	20.55	19.69	44.31	21.31	21.39

TABLE V: The word error rates (%) obtained by training PARZNETS 2D architecture using the Kaldi training fold for AMI-IHM and then decoding the validation and test folds of AMI-IHM and AMI-SDM using the learned model. We did not use i-vectors in the experiments and have trained using a cross-entropy loss function. Following the original Kaldi recipe, a 3-GRAM language model built from the AMI and FISHER data was adopted.

TRAINED ON AMI-IHM	AMI-IHM		AMI-SDM	
	DEV	EVAL	DEV	EVAL
PARZNETS 2D	25.1	26.4	76.1	85.0
PARZNETS 2D + ALGS. 1-4	26.6	29.5	52.2	59.3
MFCC & MLP [51]	-	32.3	-	76.0
MFCC & MLP + SDM OUTPUT LAYER ADAPT. [51]	-	43.7	-	57.0
MFCC & MLP + SDM INPUT LAYER ADAPT. [51]	-	44.9	-	58.1

the setting with utterance normalization and pre-emphasis the proposed approach contributes to more than 40% relative improvement in WER. This is increased to more than 50% relative if one opts not to perform utterance normalization and signal pre-emphasis. SPECAUGMENT fails to achieve comparable WER across different settings. Our hypothesis is that this is due to the fact that SPECAUGMENT has been devised to address the *memorization* effect in recurrent neural networks and, thus, it acts more as a regularizer specific to that class of models rather than providing a means of characterizing the vicinity of input/training speech signals. Our intuition is in part confirmed with the requirement to adapt the scheme for feedforward models [e.g., see 50].

E. Robustness Relative to Conversational Speech

AMI [44] is a conversational speech dataset with approximately 80 hours of speech. It comes in three parts, two of which are of interest to this work: *i*) AMI-IHM that contains data recorded using individual headset microphones, and *ii*) AMI-SDM that contains data recorded using a distant microphone. Our goal in this section is to demonstrate that the proposed augmentation schemes can be used to improve the performance of acoustic models learned on data from headset microphones in distant-talking speech recognition tasks. To this end, we first train the baseline PARZNETS 2D architecture using the training fold of Kaldi AMI-IHM recipe and evaluate its effectiveness on AMI-SDM with distant-talking speech. This is then repeated by performing *online* augmentation of training utterances using Algorithms 1-4, i.e., at the onset of each epoch one decides with the uniform probability whether to keep the original AMI-IHM training utterance or corrupt it via the proposed augmentation schemes. The alignments for this tasks were generated using the Kaldi AMI-IHM recipe configured with $3,984$ HMM state ids. The training process runs for 16 epochs with the remaining setup being the same as in other experiments.

Table V reports the results of this experiment. We observe that on both AMI-SDM test folds (labelled with DEV and EVAL) the addition of vicinal characterizations captured by Algorithms 1-4 improves the word error rate in excess of 43% relative. Our empirical result, thus, demonstrates that the proposed schemes can improve robustness of acoustic models relative to distant-talking speech. Moreover, we compare our results to prior work [51] where the goal was to improve robustness relative to distant-talking speech by means of layer adaptation using AMI-SDM data. The empirical results show that our vicinal model is competitive with layer adaptation baselines despite relying on AMI-IHM training utterances alone.

VII. DISCUSSION

The main focus of prior studies on leveraging data augmentation for improving the performance of acoustic models is on achieving gains across different benchmarks datasets. This is typically done without considering the underlying machine learning principles that are responsible for translating the additional information originating from the augmentation schemes into the robustness. This work aims to bridge this gap by posing data augmentation as an instance of vicinal risk minization. The latter is a theoretically well founded setting that allows for an insight into the inductive bias incorporated into neural networks by means for augmentation schemes. More specifically, we have demonstrated in Section III that for robustness relative to distribution shifts between train and test samples one requires a good characterization of the marginal density around training samples. Theorem 1 is our main theoretical contribution and it shows that the pre-softmax layer outputs that correspond to signals sampled from the vicinal density centered at a training sample concentrate in the feature space given by that layer of the neural architecture. As a result of this, neural networks will perform smoothing rather than interpolation and (in hybrid HMM-DNN models) assign similar likelihoods for HMM state ids across neighborhoods described by vicinal densities. We have further supported this via insights in Section III-B where it was demonstrated that learning with cross-entropy loss in the vicinal setting (i.e., with data augmentation) amounts to maximizing the log-likelihood over neighborhoods rather than individual samples. These theoretical findings were put to the test in Section VI-C where we trained a model using the Kaldi clean-condition recipe for AURORA4 and decoded the noisy/divergent test folds by employing average likelihoods relative to vicinal densities at test samples rather than standard pointwise predictions. The empirical results indicate that there is only a minor performance degradation which supports our results on inductive bias and concentration of vicinal samples. Moreover, we have evaluated

in the same manner a model trained using the Kaldi multi-condition recipe for AURORA4 and observed a significant performance degradation relative to samples from vicinal densities (see Table III).

In the second part of the paper, we focus on devising augmentation schemes that can be effective for learning models robust relative to different types and levels of stationary additive noise, as well as divergences due to microphone effects. This is followed by a focused empirical study that first characterizes the effectiveness of individual schemes relative to different sources of additive noise and microphone effects (Section VI-A). Our empirical results in this regard are compelling showing significant improvement in out-of-distribution generalization compared to training using the standard risk minimization principle. More specifically, we observe an improvement in excess of 150% relative on AURORA4 when generalizing from clean speech to noisy. This is further strengthened with empirical results on conversational speech where we show that the proposed augmentation schemes can help in the extremely difficult problem of generalizing from data collected via headset microphones to distant-talking speech. We observe an improvement in excess of 40% relative on this task (Section VI-E). We have also performed a detailed analysis of the approach relative to state-of-the-art augmentation schemes based on waveform signals and filterbank features. More specifically, the multi-condition recipe for AURORA4 offers the most effective waveform-based augmentation scheme known to us because it performs explicit matching between train and test conditions relative to types of additive noise and microphone effects. We have demonstrated that the model learned via vicinal risk minimization and the Kaldi clean-condition recipe is competitive with multi-condition training (Section VI-B), which is a difficult task to achieve. Moreover, we also show that the proposed approach offers means for learning effective models based on filterbank features (Section VI-D). In that experiment, we have contrasted our augmentation schemes to SPECAUGMENT [3] and demonstrated that we clearly outperform this baseline on the task where there is a significant divergence in acoustic properties characteristic of train and test folds. Moreover, our empirical results have identified a shortcoming of the SPECAUGMENT scheme, which we hypothesize is due to the fact that it was designed to address the hidden state memorization in recurrent neural networks. Namely, time and frequency masking forces recurrent neural networks to infer the missing information from the available context (i.e., sequence of preceding and succeeding non-masked frames). This is probably the reason why it has been coupled mainly with recurrent architectures in prior work. Thus, the proposed augmentation schemes are not only more effective but also more general and widely applicable for achieving robustness relative to additive noise and microphone effects.

VIII. CONCLUSION

We have proposed an effective approach for learning robust acoustic models and demonstrated empirically that it can generalize to unseen acoustic conditions. Our theoretical contributions show that the approach can incorporate a robust inductive bias into the learning process and that it provides a flexible method for characterizing vicinal risk estimates around training observations. The latter allows for effective out-of-distribution generalization and motivates further research in the direction of vicinal risk minimization. We have also given a bound that characterizes the robustness of waveform-based models, given in terms of the Jacobian and Hessian tensors. An interesting aspect of this bound is that it can be used as a basis for regularization mechanisms, which we aim to explore further in future work. In addition to all of this, we have also proposed highly effective data augmentation schemes and demonstrated empirically that they have the potential to address the issues with spurious correlations in acoustic models. Our ablation study carefully dissects the influence of individual schemes on the out-of-distribution generalization relative to several different noise types and microphone effects.

ACKNOWLEDGMENTS

The authors would like to thank Erfan Loweimi for providing the results of his experiments with the CLDNN baseline on AURORA4 during the revision period.

APPENDIX

Proof. We start by expanding the sufficient statistic using the Taylor theorem for multivariate functions, i.e.,

$$\begin{aligned}\Psi\left(x+\epsilon\right) = \quad & \Psi\left(x\right)+\nabla\Psi\left(x\right)^{\top}\epsilon + \\ & \frac{1}{2}\sum_{i,k=1}^{d}\nabla^{2}\Psi_{ik,*}\left(x\right)\epsilon_{i}\epsilon_{k}+o(\left\|\epsilon\right\|^{2})\ .\end{aligned}$$

From our assumptions, it follows that there exists a constant $A>0$ such that for all $x\in\mathcal{X}$ it holds $\left|\nabla^{2}\Psi_{ik,j}\left(x\right)\right|<A$, with $1\leq i,k\leq d$ and $1\leq j\leq D$. This then implies that we can upper bound the perturbation as

$$\begin{aligned}&\left\|\Psi\left(x+\epsilon\right)-\Psi\left(x\right)\right\| \leq \\ &\left\|\nabla\Psi\left(x\right)^{\top}\epsilon\right\|+\frac{1}{2}\left\|\sum_{i,k=1}^{d}\nabla^{2}\Psi_{ik,*}\left(x\right)\epsilon_{i}\epsilon_{k}\right\|+o(\left\|\epsilon\right\|^{2})\ .\end{aligned}$$

As the random variable ϵ has law $\epsilon\sim\mathcal{N}\left(0,\sigma^{2}\mathbb{I}\right)$, it follows that we can bound the third term in this inequality by $o\left(\sigma^{2}\right)$. That term also determines the approximation order for our bound and it is independent of the neighborhood centered at $x\in\mathcal{X}$. We now focus on the two leading terms and introduce a univariate random variable $Z=\left\|\nabla\Psi\left(x\right)^{\top}\epsilon\right\|+\frac{1}{2}\left\|\sum_{i,k=1}^{d}\nabla^{2}\Psi_{ik,*}\left(x\right)\epsilon_{i}\epsilon_{k}\right\|$. Then,

$$P_{\epsilon}\left(\left\|\Psi\left(x+\epsilon\right)-\Psi\left(x\right)\right\|\geq r\right)\leq P_{\epsilon}\left(Z>r\right)<\frac{\mathbb{E}_{\epsilon}\left[Z\right]}{r},$$

where the last inequality follows from the Markov bound.

Now, observe that $\nabla\Psi\left(x\right)^{\top}\epsilon$ is a Gaussian random variable that follows the distribution $\mathcal{N}\left(0,\sigma^{2}\nabla\Psi\left(x\right)^{\top}\nabla\Psi\left(x\right)\right)$. Denote $C=\nabla\Psi\left(x\right)^{\top}\nabla\Psi\left(x\right)$ and observe that this matrix is positive semi-definite. Thus, C admits an eigendecomposition $C=U\Lambda U^{\top}$ with an orthogonal matrix U and a diagonal

eigenvalue matrix Λ. This allows us to rewrite the first term from the right-hand side of the Markov bound as

$$\mathbb{E}_\epsilon \left[\left\| \nabla \Psi \left(x \right)^\top \epsilon \right\| \right] = \mathbb{E}_{v \sim \mathcal{N}(0, \sigma^2 \Lambda)} \left[\| v \| \right] \ .$$

On the other hand, from the Jensen inequality it follows that

$$\left(\mathbb{E}_v \left[\| v \| \right] \right)^2 \leq \mathbb{E}_v \left[\| v \|^2 \right] = \sigma^2 \operatorname{tr} \left(\Lambda \right) = \sigma^2 \operatorname{tr} \left(C \right) = \sigma^2 a \ ,$$

with the constant $a > 0$ introduced in Definition 1.

For the second term, observe that [52]

$$\mathbb{E}_\epsilon \left[\left\| \sum_{i,k=1}^{d} \nabla^2 \Psi_{ik,*} \left(x \right) \epsilon_i \epsilon_k \right\| \right] \leq \mu_1 \mathbb{E}_\epsilon \left[\| \epsilon \|^2 \right] \ ,$$

where μ_1 is the largest singular value of the D tensors $\nabla^2 \Psi_{**,j} \left(x \right)$ with $1 \leq j \leq D$. As ϵ is an isotropic Gaussian random variable, we have $\mathbb{E}_\epsilon \left[\| \epsilon \|^2 \right] = \sigma^2 \cdot \mathbb{E}_{u \sim \mathcal{N}(0, \mathbb{I}_d)} \left[\| u \|^2 \right] = \sigma^2 d$. While this provides a bound on the second term, there is an undesirable dependence on the dimension of the instance space d. In the following part of the proof, we show how this can be avoided by using the trace of the Hessian tensor as captured by the constant b introduced in Definition 1.

First, we observe that

$$\mathbb{E}_\epsilon \left[\left\| \sum_{i,k=1}^{d} \nabla^2 \Psi_{ik,*} \left(x \right) \epsilon_i \epsilon_k \right\|^2 \right] =$$

$$\sum_{j=1}^{D} \mathbb{E}_\epsilon \left[\left(\epsilon^\top \nabla^2 \Psi_{**,j} \left(x \right) \epsilon \right)^2 \right] = \sum_{j=1}^{D} \mathbb{E}_{v \sim \mathcal{N}[0, \sigma^2 \mathbb{I}]} \left(v^\top \Xi_j v \right)^2 =$$

$$\sum_{j=1}^{D} \mathbb{E}_{v \sim \mathcal{N}(0, \sigma^2 \mathbb{I})} \left[\left(\sum_{k=1}^{d} \xi_{jk} v_k^2 \right)^2 \right] =$$

$$2\sigma^4 \sum_{j=1}^{D} \operatorname{tr} \left(\nabla^2 \Psi_{**,j} \left(x \right) \nabla^2 \Psi_{**,j} \left(x \right) \right) + \operatorname{tr} \left(\nabla^2 \Psi_{**,j} \left(x \right) \right)^2 \ ,$$

where Ξ_j is a diagonal eigenvalue matrix with entries ξ_{jk} and $1 \leq k \leq d$. The latter eigenvalues are from the decomposition of the symmetric matrix $\nabla^2 \Psi_{**,j} \left(x \right) \in \mathbb{R}^{d \times d}$. Now, combining the latter with the Jensen inequality we have the following upper bound on the second order term

$$\mathbb{E}_\epsilon \left[\left\| \sum_{i,k=1}^{d} \nabla^2 \Psi_{ik,*} \left(x \right) \epsilon_i \epsilon_k \right\| \right] \leq \sigma^2 \sqrt{2b} \ .$$

Hence, we have that it holds

$$\mathbb{E}_\epsilon \left[Z \right] < \sigma \left(\sqrt{a} + \sigma \sqrt{b/2} \right) \ . \tag{6}$$

This then implies that for all $\delta > 0$ and $r > \frac{\sigma}{\delta} \left(\sqrt{a} + \sigma \sqrt{b/2} \right)$

$$P_\epsilon \left(\| \Psi \left(x + \epsilon \right) - \Psi \left(x \right) \| \geq r \right) < \delta \ .$$

□

Remark. The result in Theorem 1 applies within the second order Taylor expansion, due to the extra term that was omitted from consideration and which can be bounded by $o\left(\sigma^2\right)$.

REFERENCES

[1] D. Yu, M. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks - studies on speech recognition," in *International Conference on Learning Representations*, 2013.

[2] E. Vincent, S. Watanabe, A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, vol. 46, pp. 535–557, 2017.

[3] D. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. Cubuk, and Q. Le, "Specaugment: A simple augmentation method for automatic speech recognition," in *INTERSPEECH*, 2019.

[4] K. Kumar, B. Ren, Y. Gong, and J. Wu, "Bandpass noise generation and augmentation for unified ASR," in *INTERSPEECH*, 2020.

[5] M. Ravanelli and Y. Bengio, "Speech and speaker recognition from raw waveform with SincNet," *arXiv:1812.05920*, 2018.

[6] E. Loweimi, P. Bell, and S. Renals, "On learning interpretable CNNs with parametric modulated kernel-based filters," in *INTERSPEECH*, 2019.

[7] D. Oglic, Z. Cvetkovic, P. Bell, and S. Renals, "A deep 2D convolutional network for waveform-based speech recognition," in *INTERSPEECH*, 2020.

[8] L. Alsteris and K. Paliwal, "Further intelligibility results from human listening tests using the short-time phase spectrum," *Speech Communication*, vol. 48, 2006.

[9] B. Meyer, M. Wächter, T. Brand, and B. Kollmeier, "Phoneme confusions in human and automatic speech recognition," *INTERSPEECH*, 2007.

[10] S. Peters, P. Stubley, and J.-M. Valin, "On the limits of speech recognition in noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1999.

[11] M. Ager, Z. Cvetkovic, and P. Sollich, "Combined waveform-cepstral representation for robust speech recognition," *IEEE ISIT*, 2011.

[12] J. Yousafzai, P. Sollich, Z. Cvetkovic, and B. Yu, "Combined features and kernel design for noise robust phoneme classification using support vector machines," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 1396–1407, 2011.

[13] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[14] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1469–1477, 2015.

[15] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[16] N. Joy, D. Oglic, Z. Cvetkovic, P. Bell, and S. Renals, "Deep scattering power spectrum features for robust speech recognition," in *INTERSPEECH*, 2020.

[17] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, 1962.

[18] V. Epanechnikov, "Non-parametric estimation of a multivariate probability density," *Theory of Probability and its Applications*, vol. 14, pp. 153–158, 1969.

[19] J. Ba, R. Kiros, and G. Hinton, "Layer normalization," *arXiv:1607.06450*, 2016.

[20] E. Jaynes, "Information theory and statistical mechanics," *Physical Review*, vol. 106, pp. 620–630, 1957.

[21] Y. Altun, A. Smola, and T. Hofmann, "Exponential families for conditional random fields," in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 2004.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[23] O. Chapelle, J. Weston, L. Bottou, and V. Vapnik, "Vicinal risk minimization," in *Advances in Neural Information Processing Systems*. MIT Press, 2001.

[24] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 2019, pp. 1310–1320.

[25] H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang, "Provably robust deep learning via adversarially trained smoothed classifiers," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[26] T. Dao, A. Gu, A. Ratner, V. Smith, C. De Sa, and C. Re, "A kernel theory of modern data augmentation," in *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 2019, pp. 1528–1537.

[27] N. Parihar and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," 2002.

[28] C. Wang and W. Xiao, "Second-order IIR notch filter design and implementation of digital signal processing system," in *Instruments, Measurement, Electronics and Information Engineering*, vol. 347. Trans Tech Publications Ltd, 2013.

[29] T. Ko, V. Peddinti, D. Povey, M. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *INTERSPEECH*, 2017.

[30] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A Python package for audio room simulations and array processing algorithms," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[31] L. Drude, J. Heitkaemper, C. Boeddeker, and R. Haeb-Umbach, "SMS-WSJ: Database, performance measures, and baseline recipe for multi-channel source separation and recognition," *arXiv:1910.13934*, 2019.

[32] R. Lippmann, E. Martin, and D. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1987.

[33] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Ng, "Deep speech: Scaling up end-to-end speech recognition," *arXiv pre-print: 1412.5567*, 2014.

[34] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *INTERSPEECH*, 2015.

[35] N. Jaitly and G. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proceedings of the International Conference on Machine Learning*, 2013.

[36] D. Oglic, Z. Cvetkovic, and P. Sollich, "Learning waveform-based acoustic models using deep variational convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[37] A. Laptev, R. Korostik, A. Svischev, A. Andrusenko, I. Medennikov, and S. Rybin, "You do not need more data: Improving end-to-end speech recognition by text-to-speech data augmentation," in *Proceedings of the 13th Int. Congress on Image and Signal Processing, Bio-medical Eng. and Inf.* IEEE, 2020.

[38] A. Renduchintala, S. Ding, M. Wiesner, and S. Watanabe, "Multi-modal data augmentation for end-to-end ASR," *arXiv pre-print: 1803.10299*, 2018.

[39] A. Ragni, K. Knill, S. Rath, and M. Gales, "Data augmentation for low resource languages," in *INTERSPEECH*, 2014.

[40] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv pre-print: 1904.05862*, 2019.

[41] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv pre-print: 2006.11477*, 2020.

[42] L. Wang, M. Gales, and P. Woodland, "Unsupervised training for mandarin broadcast news and conversation transcription," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[43] L. Wang and P. Woodland, "Discriminative adaptive training using the MPE criterion," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003.

[44] S. Renals, T. Hain, and H. Bourlard, "Recognition and interpretation of meetings: AMI and AMIDA," in *IEEE ASRU*, 2007.

[45] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE ASRU*, 2011.

[46] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2010.

[47] T. Sainath, R. J. Weiss, K. Wilson, A. W. Senior, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *INTERSPEECH*, 2015.

[48] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2016.

[49] J. Rownicka, P. Bell, and S. Renals, "Multi-scale octave convolutions for robust speech recognition," in *IEEE ASRU*, 2019.

[50] X. Li, Y. Zhang, X. Zhuang, and D. Liu, "Frame-level SpecAugment for deep convolutional neural networks hybrid ASR systems," *IEEE SLT*, 2021.

[51] I. Himawan, P. Motlicek, D. Imseng, B. Potard, N. Kim, and J. Lee, "Learning feature mapping using deep neural network bottleneck features for distant large vocabulary speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4540–4544.

[52] S. Lang, *Linear algebra*, 3rd ed., ser. Undergraduate Texts in Mathematics. New York: Springer-Verlag, 1987.