TOWARDS LEARNING NUISANCE-FREE REPRESENTATIONS OF SPEECH

Lixing Liu, Sayan Ghosh, Stefan Scherer

Institute for Creative Technologies, Department of Computer Science
University of Southern California, Los Angeles, CA, USA
{lxliu, sghosh, scherer}@ict.usc.edu

ABSTRACT

Representation learning methods, such as deep autoencoders, have received sustained attention due to their ability to effectively learn meaningful representations for a variety of applications. While these learning approaches are able to derive representations from any source signal (e.g., images, language, or voice signals) and encourage the separation in dominating factor domains, they broadly treat factors of variation pertaining to nuisances (e.g., recording conditions, gender of speaker, accent etc.) no different from often subtle more interesting factors, such as paralinguistic target variables (e.g., voice quality and phonetic vowels). In paralinguistic speech analyses, nuisance variables (e.g. gender and accent of speakers) often dominate acoustic subtleties that pertain for example to the affect or well-being of the speaker. In this work, we seek to capture nuisance-free embeddings by learning two separate orthogonal representations: one representation specialized to capture nuisance factors and one that improves the representation of the target. We propose unsupervised and (semi-)supervised orthogonal autoencoders that allow us to learn informative representations of paralinguistic and phonetic targets while removing the effect of the nuisance – gender. Overall, our proposed model outperforms state-of-the-art approaches and shows improved target representations.

Index Terms— Computational paralinguistic, Nuisance-free learning, Representation learning, Affective computing

1. INTRODUCTION

Deep neural networks have made significant strides in a wide variety of learning tasks, setting new performance benchmarks in audio-visual recognition and natural language understanding. Machine learning, in general, and automatic human behavior understanding, in particular, rely on good data representations that have a good discriminatory faculty in classification and regression experiments, such as emotion recognition from speech [1, 2].

To derive efficient representations of data, researchers have adopted two main strategies: (a) carefully crafted and tailored feature extractors designed for particular tasks [3] and (b) algorithms that learn representations automatically from the data [4]. Previous research in affective speech recognition mainly focused on using off-the-self feature extractors, such as Mel-Frequency Cepstral Coefficients (MFCCs), fundamental frequency, or features pertaining to the glottal flow dynamics, including normalized amplitude quotient or quasi open quotient [5]. While representation learning on spectrograms has increasingly become the new standard for human behavior understanding and recognition [6], there has been less progress in exploring representations for nonverbal or paralinguistic attributes from speech.

Voice quality, which refers to the coloring or timbre of the voice, for example, is closely related to the speaker's affect or emotion [7]. Earlier work has also investigated its implications for applications such as mental status evaluation and assessment of psychological distress [8, 9]. When learning representations of a speaker's voice quality, it might be beneficial for the model to remove the effects of the speaker's gender from the signal, as a number of voice characteristics are strongly dependent on gender due to anatomical differences [10]. Gender and other factors can be considered a nuisance when attempting to learn representations of affective speech or voice quality. Within this work, we seek to learn representations that are independent of factors of variation that pertain to such nuisance variables. We hypothesize that by removing effects of gender we can improve representations of voice quality. Specifically, we seek to learn representations of three voice quality categories: *Tense*, *Modal* and *Breathy* voice. As an additional task, we seek to learn representations of vowels independent of gender to further support our hypothesis.

Our work is based on the autoencoder architecture [11]. Autoencoders are neural networks typically trained to learn a lower-dimensional distributed representation of the input data. This is one of the most common architectures in learning representations of a range of signals, including acoustic inputs [2]. In this work, we primarily focus on investigating the separation of paralinguistic target factors and the labeled nuisance factors in the latent spaces of autoencoders. To ensure that part of the learned latent embedding represents the target, we enforce a separation between the latent space representing the target variable (i.e., voice quality and vowels) and

the latent space representing the nuisance variable (i.e., gender) through an additional orthogonality loss. Specifically, we seek to minimize the cosine similarity of the target space and the nuisance space in an additional regularization term in the learning objective to achieve orthogonality.

Within this work we seek to investigate a number of objectives. Specifically, we hypothesize that it is possible to learn representations that are informative for two target variables (a) voice quality and (b) phonetic vowels, while removing effects of nuisance variables (i.e., gender of speaker). We further hypothesize that high-level representations of target variables can be learned using our proposed orthogonal autoencoders in (a) unsupervised, (b) supervised, and (c) semi-supervised fashion.

2. RELATED WORK

The deep autoencoder has been widely used in acoustic signal analyses. For example, Feng et al. [12] investigated noisy reverberant speech recognition using denoising autoencoder. More recently, Deng et al. [13] developed semi-supervised autoencoders for speech emotion recognition.

In order to reconstruct the input signal, the latent representation within such autoencoders must maintain most factors of variation in the input. Moreover, in order to be able to disentangle certain factors of variation, such as speaking style, speaker gender, etc., we need to develop more complex models. Typically, additional subnets and correlated objectives, which are often jointly optimized, need to be incorporated in more complex networks. One such approach to disentangle writing style and content on MNIST handwritten digits was introduced by Cheung et al. [14] using a cross-covariance penalty. However, speech can be much more complex than highly constrained tasks, such as handwritten digits. Hence, the undesired variables (e.g., gender of speaker, microphone distortions) are still entangled with other factors across the latent variables if not explicitly or implicitly imposed in the learning objective.

Louizos et al. [15] introduced the variational fair autoencoder by combining a basic variational autoencoder [16] with additional regularization terms to remove certain nuisance variables. The model successfully learned representations that are devoid of undesired factors, while retaining as much information as possible in the latent space. This model was again trained and tested on the constrained MNIST datasets. In parallel, Makhzani et al. [17] developed a similar approach, namely the adversarial autoencoder, which uses an adversarial training procedure for variational inference by matching the aggregated posterior of hidden vector with the prior distribution. Bousmalis et al. [18] further developed a private-shared encoder framework for learning domain-invariant image representations of both source and target domains, which captures specific properties of both domains in object classification and pose estimation tasks.

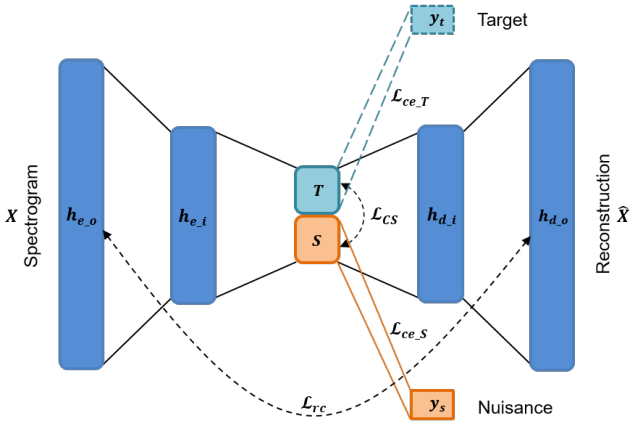

Fig. 1. Orthogonal autoencoder architecture. The softmax prediction layers y_s and y_t, which only applied for (semi-) supervised tasks, are incorporated in the signal reconstruction process. The sizes of the hidden layers are always {512, 256, 64+64} with 774-dimension input spectrograms.

While previous approaches to augment autoencoders for factor disentangling are mainly focused on image datasets, we attempt to leverage these technologies to model acoustic signals directly and disentangle nuisance factors from paralinguistic targets. Further, we attempt to complement the previously proposed works with an additional model that is well suited to learn nuisance-free representations of speech.

3. PROPOSED MODEL

As mentioned earlier, our proposed model – the orthogonal autoencoder (OAE) – is based on the basic autoencoder architecture. To learn improved representations of the target and nuisance variables, we propose to add an additional orthogonality penalty into the autoencoder objective to encourage the independence of target and nuisance representations. Overall, the loss function to train the orthogonal autoencoder consists of three separate objectives: (1) orthogonality between target and nuisance space, (2) reconstruction loss, which is typical for autoencoders, and (3) classification losses for recognition of nuisance variable using the nuisance space representation S and additionally in the (semi-)supervised case the recognition of the target variable using the target space representation T. We provide details for each of these objectives below.

Orthogonality. In order to reduce the correlation between target space T and nuisance space S, we introduce a cosine similarity loss $\mathcal{L}_{cs}$ (1) to enforce orthogonality.

$$\mathcal{L}_{cs} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{t}_i \mathbf{s}_i \, ||\mathbf{t}_i||_2^{-1} \, ||\mathbf{s}_i||_2^{-1} \tag{1}$$

where N is the size of input samples. $\mathbf{t}$ and $\mathbf{s}$ are latent space representations of the target and the nuisance, respectively.

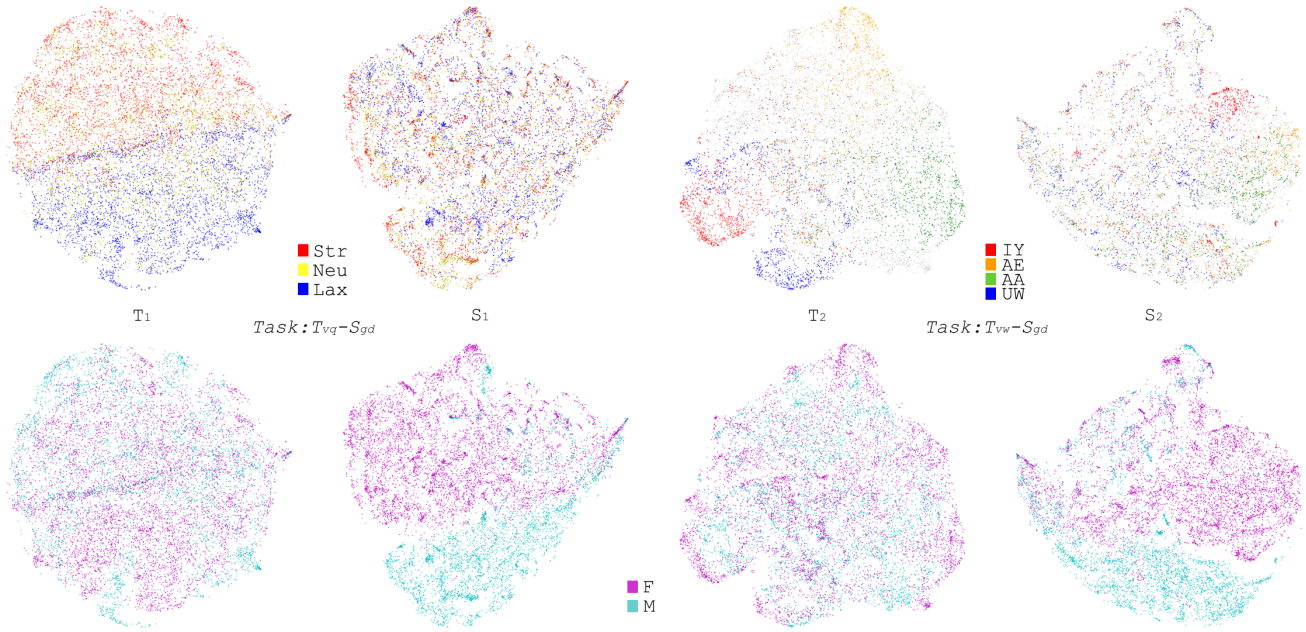

Fig. 2. t-SNE embedding visualization of phonetic-level representations of voice quality (Red- *Str/Tense*; Yellow - *Neu/Modal*; Blue - *Lax/Breathy*), gender (Magenta - Female; Cyan - Male) and vowel (Red - IY; Orange - AE; Green - AA; Blue - UW; Grey - Other 5 vowels). The left side refers to the voice quality classification task while disentangling gender. The right side denotes the phonetic vowel classification task while removing gender. Both tasks are using proposed supervised OAE model.

Task	Target T_{vq} vs. Nuisance S_{gd}				Target T_{vw} vs. Nuisance S_{gd}			
Variable	vq(3)		gd(2)		vw(9)		gd(2)	
NLP on raw spectrogram	51.0520		92.4917		41.3237		92.4917	
AE	49.6247		92.6607		41.0456		92.6607	
VAE [16]	48.4427		92.0467		32.4805		92.0467	
Latent Space	T_1		S_1		T_2		S_2	
Variable	**vq (3)**	gd (2)	vq (3)	**gd (2)**	**vw (9)**	gd (2)	vw (9)	**gd (2)**
VFAE (unsupervised) [15]	49.4022	83.4584	50.2225	94.6886	35.1780	87.8232	33.7319	94.5217
DSN (supervised) [18]	59.3437	88.2091	55.5741	94.6242	48.3037	85.9010	40.5451	94.1046
OAE (unsupervised)	49.5505	85.8454	52.1413	94.6607	36.9197	86.0112	35.1780	94.8276
OAE (semi-supervised; 5%)	56.2413	88.3204	50.5143	93.9672	44.6205	86.3775	34.7234	94.1061
OAE (semi-supervised; 40%)	59.4587	87.2937	52.3081	94.1324	48.4267	84.8763	36.2246	**94.9300**
OAE (supervised)	**62.8476**	88.2680	57.2997	**94.7442**	**54.4721**	85.1780	39.9889	94.3827

Table 1. Evaluation metrics for models on different target-nuisance settings, averaged across folds. Unweighed accuracies in % for 3-class voice quality (vq (3)), 9-class vowels (vw (9)), as well as 2-class gender (gd (2)) classifications are reported.

Reconstruction. The orthogonality loss is further accompanied by the typical reconstruction loss $\mathcal{L}_{rc}$ (2) formed by the batch-normalized inputs $\mathbf{x}$ and the corresponding outputs $\hat{\mathbf{x}}$ of the autoencoder as follow:

$$\mathcal{L}_{rc} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \log \hat{\mathbf{x}}_i + (\mathbf{1} - \mathbf{x}_i) \log (\mathbf{1} - \hat{\mathbf{x}}_i) \quad (2)$$

Classification. We further minimize classification losses $\mathcal{L}_{ce\_S}$ and $\mathcal{L}_{ce\_T}$ (3). For both (semi-)supervised and unsupervised cases, we minimize the cross-entropy cost of nuisance variable classification on labeled mini-batches. The negative log-likelihood loss for the nuisance embedding and the loss of target embedding in our supervised batches are:

$$\mathcal{L}_{ce\_S} = -\sum_{i=1}^{N} \mathbf{y}_i^s \log \hat{\mathbf{y}}_i^s; \;\; \mathcal{L}_{ce\_T} = -\sum_{i=1}^{N} \mathbf{y}_i^t \log \hat{\mathbf{y}}_i^t \quad (3)$$

where $\mathbf{y}$ is the one-hot label and $\hat{\mathbf{y}}$ is the softmax prediction. Hence, the overall training objective (4) is to minimize the

linear combination of the losses above and the final objective functions of unsupervised $\mathcal{L}_u$ and supervised $\mathcal{L}_s$ steps are:

$$\begin{aligned} \mathcal{L}_u &= \mathcal{L}_{rc} + \alpha\mathcal{L}_{cs} + \beta\mathcal{L}_{ce\_S} \\ \mathcal{L}_s &= \mathcal{L}_{rc} + \alpha\mathcal{L}_{cs} + \beta_S\mathcal{L}_{ce\_S} + \beta_T\mathcal{L}_{ce\_T} \end{aligned} \tag{4}$$

For all (semi-)supervised and unsupervised tasks, both encoders and decoders have two middle hidden layers. The inner layers are half the size (256 units) of the outer ones (512 units). To ensure that the bottle-neck embeddings are able to capture sufficient information, the latent spaces, T and S, both contain 64 units. As shown in Figure 1, T and S are concatenated to reconstruct the signal. While, the softmax prediction layers are just connected to one specific latent space.

Baseline Models. As our baseline models for the experiments we deployed: (0) softmax on raw spectrogram, (1) vanilla autoencoders (AE), (2) variational autoencoders (VAE) [16], (3) unsupervised variational fair autoencoders (VFAE) [15], and (4) supervised domain separation networks (DSN) [18] with our cosine orthogonality term. The network architectures are all the same size as the proposed model.

4. DATA AND EXPERIMENTAL SETUP

An enlarged version of the Cereproc acted speech dataset [19] was leveraged for all experiments. Cereproc granted us access to an internal dataset, which contains 76,427 partially labeled utterances. The raw waveform (16kHz) was further processed through spectrogram extraction with 129 Fast Fourier Transform (FFT) bins at 0.08 seconds per frame.

We manipulate phonetic-level segmentations to learn representations of vowels and voice quality. According to the ARPAbet phonetic transcription symbols in the annotations, we only use the monophthong sounds since these pure vowel sounds carry out the most clear patterns defined by their utterance-level voice quality labels. For our experiments, monophthong UH is excluded due to its underrepresentation. We utilize segments of 9 phonemes (AO, AA, IY, UW, EH, IH, AH, AX and AE) with a fixed length shifting window of 0.48 seconds (6 spectrogram frames) so that the input vector of the autoencoders is 774 dimensions. Sample windows are centered at the annotated apex time stamp for each phoneme.

Approximately 60,000 gender-balanced utterance samples with meta data were processed and validated. A subset, 17,982 samples, with fully balanced labels (voice quality and phonetic vowel) spoken by 5 speakers in Received Pronunciation (RP) English has been chosen for the (semi-)supervised approaches. To render our experiments as rigorous as possible and to test generalizability of our approach, we conducted leave-out testing and randomized cross-validation experiments with disjoint training, testing, and validation sets.

Parameter fine-tuning of α, β_S, β_T in the objective (4) has been conducted on $\{0.1, ..., 1.0\}$ in steps of 0.1 and $\{5.0, ..., 255.0\}$ in steps of 5.0. Then, the best parameter set was adjusted in steps of 0.5. In terms of target classification accuracies for supervised tasks, the best performance is achieved when $\alpha = 5.0, \beta_S = 10.0$ and $\beta_T = 75.0$ for voice quality and $\alpha = 9.5, \beta_S = 10.5$ and $\beta_T = 70.5$ for vowels, with the batch size of 132.

5. RESULTS AND CONCLUSIONS

In Table 1, we report performance of the state-of-the-art baseline models and the here proposed model (see Section 3). It can be seen that for both tasks, namely voice quality and vowel recognition, the proposed OAE model performs best and outperforms baseline models with respect to nuisance-variable gender and paralinguistic target variables.

As expected, gender is the dominating factor within the representations and is classified with accuracies north of 90%. While for both vowels (linguistic content) and voice quality tasks, OAE performs the best. It is clear that some supervision is required to achieve good recognition results. Both target variables are best recognized using the fully supervised approach, however, performance improves significantly with minimal guidance, through a small portion of target labels. With 5% labeled batches, the semi-supervised experiment is able to achieve a better voice quality classification performance in T space than that is in S. Supervision of 40% using the semi-supervised network reaches similar target accuracies to the fully supervised baseline model.

Figure 2 shows vowel-level scatter plots of 10,000 labeled samples using 2D t-SNE [20] with 1,000 iterations. The first two columns denote target space (T) and nuisance space (S) of supervised voice quality recognition task, respectively. The right two columns refer the T and S of vowel recognition task. The nuisance-free target representations with clear separations between *Lax* and *Str* voice quality as well as the four cardinal vowels can be explicitly seen in the T spaces. Gender on the other hand is only well separated in the nuisance space S while scattered throughout the target space T.

While preliminary, we believe the presented work is an important step towards nuisance-free representations of paralinguistic factors of speech. The introduced orthogonal autoencoder (OAE) helps generate decent target and nuisance representations simultaneously. With fewer obstacles to data collection and use, the semi-supervised approach shows its potential for paralinguistic representation learning from acoustic signals while keeping annotation costs at a minimum.

Acknowledgements. The authors would like to thank the Cereproc Inc. and in particular Matthew Aylett for sharing the data and the inspiring discussions. This material is based upon work supported by the U.S. Army Research Laboratory under contract number W911NF-14-D-0005. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Government.

6. REFERENCES

[1] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5200–5204.

[2] Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer, "Representation learning for speech emotion recognition.," in *Interspeech*, 2016, pp. 3603–3607.

[3] Piotr Dollár, Zhuowen Tu, Hai Tao, and Serge Belongie, "Feature mining for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[5] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer, "Covarepa collaborative voice analysis repository for speech technologies," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 960–964.

[6] Navdeep Jaitly and Geoffrey E Hinton, "A new way to learn acoustic events," *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[7] Christer Gobl, Ailbhe Ni, et al., "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1, pp. 189–212, 2003.

[8] Stefan Scherer, Giota Stratou, Jonathan Gratch, and Louis-Philippe Morency, "Investigating voice quality as a speaker-independent indicator of depression and ptsd.," in *Interspeech*, 2013, pp. 847–851.

[9] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[10] BjöRn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian MüLler, and Shrikanth Narayanan, "Paralinguistics in speech and language? state-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[11] Pierre Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 2012, pp. 37–49.

[12] Xue Feng, Yaodong Zhang, and James Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1759–1763.

[13] Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Frühholz, and Björn Schuller, "Semisupervised autoencoders for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 31–43, 2018.

[14] Brian Cheung, Jesse A Livezey, Arjun K Bansal, and Bruno A Olshausen, "Discovering hidden factors of variation in deep networks," *arXiv preprint arXiv:1412.6583*, 2014.

[15] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel, "The variational fair autoencoder," *arXiv preprint arXiv:1511.00830*, 2015.

[16] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[17] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[18] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan, "Domain separation networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 343–351.

[19] Matthew P Aylett and Christopher J Pidcock, "The cerevoice characterful speech synthesiser sdk," in *International Conference on Intelligent Virtual Agents (IVA)*, 2007, pp. 413–414.

[20] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.