INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Multiview Shared Subspace Learning across Speakers and Speech Commands

Krishna Somandepalli, Naveen Kumar, Arindam Jati, Panayiotis Georgiou, Shrikanth Narayanan

Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA

somandep@usc.edu, shri@ee.usc.edu

Abstract

In many speech processing applications, the objective is to model different modes of variability to obtain robust speech features. In this paper, we learn speech representations in a multiview paradigm by constraining the views to known modes of variability such as speakers or spoken words. We use deep multiset canonical correlation (dMCCA) because it can model more than two views in parallel to learn a shared subspace across them. In order to model thousands of views (e.g., speakers), we demonstrate that stochastically sampling a small number of views generalizes dMCCA to the larger set of views. To evaluate our approach, we study two different aspects of the Speech Commands Dataset: variability among the speakers and speech commands. We show that, by treating observations from one mode of variability as multiple parallel views, we can learn representations that are discriminative to the other mode. We first consider different speakers as views of the same word to learn their shared subspace to represent an utterance. We then constrain the different words spoken by the same person as multiple views to learn speaker representations. Using classification and unsupervised clustering, we evaluate the efficacy of multiview representations to identify speech commands and speakers.

Index Terms: multiview learning, speech commands, multiset canonical correlation analysis

1. Introduction

Consider a person interacting with a conversational virtual agent. A key research problem here is to recognize a speech command (e.g., "Okay Google, TV On") irrespective of who the speaker is or what device one is using. This is a challenging task because of several sources of variability in speech audio such as speaker or channel characteristics and background noise. A similar research problem, in computer vision is to identify an object imaged under different illumination and camera angles. In such problems, the different modes of variability are akin to multiple observations or *views* of the underlying phenomenon or the *signal*. Our objective is to learn a subspace from these multiple views to capture the information shared across them.

Multiview learning has been studied in several fields (See [1] for a survey). Two important research questions here are: 1) How do we handle multiple views simultaneously? and 2) How can we model a large (possibly thousands) number of views? In applications with only two views, Canonical correlation analysis (CCA [2]), and its extension, deep CCA [3] have been successfully used. In order to model more than two views in parallel, deep multiset CCA (dMCCA [4]) has been proposed. This method which is based on MCCA [5, 6] models both the between-view and within-view variance to learn a shared subspace using an independent deep neural network (DNN) per view. However, it is unclear if dMCCA can model a large number of views in parallel, since the number of parameters to be trained scale with the number of views in the dataset.

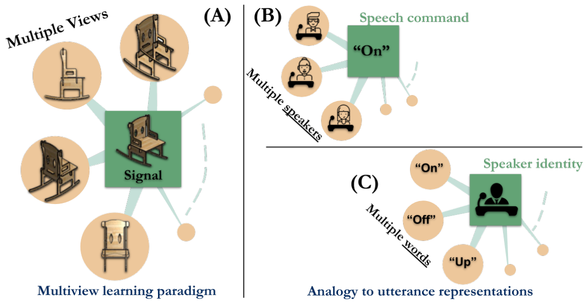

Figure 1: *(A) Representing the 'signal' from multiple 'views' (B) speakers as views (C) words as views*

In this paper, we propose a novel direction to model speech in a multiview paradigm, where the variability across speakers or channels can be considered as multiple views. Figure 1 illustrates the analogies of multiview learning for speech tasks. The acoustic features of an utterance e.g., "On" by different speakers is analogous to imaging an object from different angles. We can then constrain our views to either the speakers or the utterances to capture the shared information, i.e., the *signal*.

Related work in speech research has mostly applied multiview learning in a *multimodal* setup. Typically, video [7] or articulatory measurements [8, 9] are available along with speech. In contrast, for speech applications, we only need acoustic features, where we consider the different modes of variability as multiple views. Additionally, we propose to stochastically sample a small number of views during training to apply dMCCA for a larger set of views. For this purpose, we conduct experiments on the Speech Commands Dataset [10]. This publicly available corpus provides an excellent test-bed for us because it consists of over 1800 speakers with one or more of thirty speech commands each, thus providing a large number of views to test our approach. Specifically, we test our approach for two tasks as illustrated in Fig. 1B–C: command and speaker identification.

Because we introduce our approach within the framework of speaker and command identification, we also explore existing methods for these tasks. Obtaining utterance representations robust to speaker or channel variability has been extensively studied with total variability modeling (TVM) [11]. It is shown to perform effectively for speaker and language identification [12, 13]. TVM estimates a single low-rank representation (called *i-vector*) for all sources of variability in the data. As such, the dominant modes of variability captured largely depend on the training data, and cannot be explicitly supervised. Because they are trained in an unsupervised manner, linear discriminant analysis (LDA) [14] or probabilistic LDA (PLDA [15]) is applied to the i-vectors prior to an identification task. Alternatively, DNNs can be trained end-to-end for tasks such as speaker verification in a supervised fashion [16, 17]. To make use of back-end technology developed for i-vectors, a two-part

http://dx.doi.org/10.21437/Interspeech.2019-3130

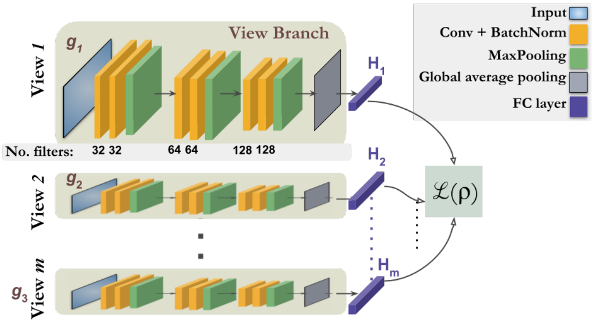

Figure 2: *CNN architecture for the view branches in dMCCA*

system: DNN to learn embeddings (x-vectors), and a separate classifier trained for the task was proposed in [18]. More recently, this method has been further improved using data augmentation [19] for robust speaker recognition.

Our primary contribution in this work is an approach to model a much larger number of views for multiview learning by stochastically sampling a smaller number of views during training within the framework of dMCCA. While dMCCA can be applied for any DNN architecture, in this paper, we use convolutional neural networks (CNN) to obtain shared representations of the speech features with the utterances considering different speakers or words as views to classify the words or speakers respectively.

2. Generalized Deep Multiset CCA

Consider N samples of D-dimensional features observing the same underlying phenomenon (signal) from M different views. Let $\mathbf{X}_l \in \mathbb{R}^{D \times N} l = 1, ..., M$, be the data matrix for the l-th view with columns as features. We define *multi-view correlation* matrix $\mathbf{\Lambda}$ as the normalized ratio of the sum of between-view covariances $\mathbf{R}_B$ and sum of within-view covariances $\mathbf{R}_W$ for M views as follows:

$$\mathbf{\Lambda} = \frac{1}{M-1}\frac{\mathbf{R}_B}{\mathbf{R}_W} = \frac{\sum_{l=1}^{M}\sum_{k=1,k\neq l}^{M} \bar{\mathbf{X}}_l(\bar{\mathbf{X}}_k)^\top}{(M-1)\sum_{l=1}^{M}\bar{\mathbf{X}}_l(\bar{\mathbf{X}}_l)^\top} \tag{1}$$

where $\bar{\mathbf{X}}_* = \mathbf{X}_* - \mathbb{E}(\mathbf{X}_*)$ are mean centered data matrices. The common scaling factor $(N-1)^{-1}M^{-1}$ in the ratio is omitted. In related work, the ratio in Eqn. 1 is maximized for MCCA [5, 20]. A variation of this measure called intraclass correlation coefficient [21] has been extensively used to quantify repeatability of measurements in test-retest studies.

Our objective is to estimate a shared subspace, $\mathbf{V} \in \mathbb{R}^{D \times D}$ such that the multi-view correlation is maximized. Formally:

$$\rho^{(M)} = \max_{\mathbf{V}} \frac{1}{D(M-1)} \frac{\text{tr}(\mathbf{V}^\top \mathbf{R}_B \mathbf{V})}{\text{tr}(\mathbf{V}^\top \mathbf{R}_W \mathbf{V})} \tag{2}$$

where tr$(\cdot)$ denotes the trace of a matrix. Note that ρ is bounded above by 1. The subspace $\mathbf{V}$ can be estimated by solving the generalized eigenvalue problem to simultaneously diagonalize $\mathbf{R}_B$ and $\mathbf{R}_W$. Hence, our objective function is the average of the ratio of eigenvalues of between-view and within-view covariances. In other words, if $\mathbf{R}_W$ is invertible then we want to find the a subspace direction that maximizes the spectral norm of $\mathbf{R}_W^{-1}\mathbf{R}_B$, thus accounting for the different modalities of variance (within-view) for a given signal-of-interest (between-view). We then compute the multi-view representation $\mathbf{Y}_l = \mathbf{V}^\top \mathbf{X}_l$ where $\mathbf{Y}_l \in \mathbb{R}^{D \times N}$ is a signal that is maximally correlated across views.

Table 1: *Speaker and utterance (utt.) characteristics*

Distribution	(Min, Max)	Mean	SD
Samples per utt.	(1484, 2203)	1941.7	287.6
Utt. per speaker	(1, 205)	31.2	25.6
	No. samples (No. speakers)		
Task	Train	Dev	Test
Command-ID	35067 (1123)	9137 (300)	14048 (445)
Speaker-ID	27314 (1428)	2092 (146)	2098 (146)

To extend MCCA across large and complex datasets, deep MCCA has been proposed [4] where each view of the data is input to a neural network $g_l(\cdot), l = 1, ..., M$. The multiview representation, in this case is the output of the last layer $\mathbf{H}_l^K = g_l(\mathbf{X}_l)$. The covariance matrices, $\mathbf{R}_B$ and $\mathbf{R}_W$ in Eq. 1 are estimated using $\mathbf{H}_l$ instead of $\mathbf{X}_l$ to maximize $\rho^{(M)}$ (Eq. 2).

In dMCCA, data from multiple views is modeled with identical and independent neural networks that are trained jointly to maximize ρ. We refer to these individual networks as *view-branches*. For speech processing applications, where we often deal with a large number of speakers the view-branches would also be in the order of thousands. In this case, dMCCA cannot be applied directly for several reasons: 1) the number of parameters increase by order of M, and may not be computationally feasible to train, 2) limited amount of data compared to the increased number of parameters 3) it is computationally expensive to estimate the covariance matrices $\mathbf{R}_B, \mathbf{R}_W$, and 4) may overfit because of high variance from modeling all the views (speakers) simultaneously.

In order to address these limitations, we propose a stochastic approach where a small number of views $m << M$ are uniformly sampled from the larger set of views. Formally, the modified objective is:

$$\rho^* = \mathbb{E}_{m \sim \mathcal{U}(1,M)}\rho^{(m)} \approx \rho^{(M)} \tag{3}$$

We refer to this method as generalized dMCCA, since it can potentially model a much larger number of views. The training procedure is weakly supervised. For example, in a classification problem, we only have the knowledge that certain samples measure the same event from different views e.g., same word "On" said by different speakers. The number of classes, or the number of samples per class is not known during training.

Another benefit of this approach is that it is *view-agnostic* because during training, we do not keep track of the specific views. We randomly pick different views for the different branches. Thus, during inference we only use one of the m trained view-branches, to obtain the representations. Because we learn a *shared* subspace across the views, we expect the representations from different branches to be maximally correlated as shown in [4].

3. Experiments

As described in Sec. 1, we wish to study two distinct aspects of variability in Speech Commands Dataset [10] in a multiview setup, by constraining one mode as view to learn discriminative representations for the other mode. For this purpose we setup two experiments: 1) **command-ID**: speakers as views to discriminate between words, and 2) **speaker-ID**: different words

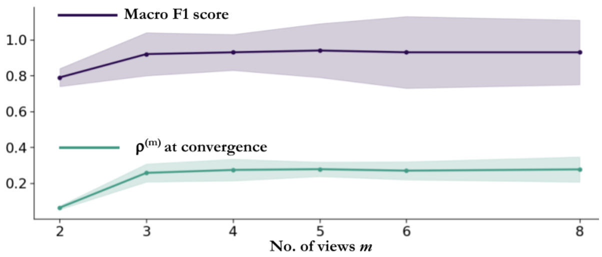

Figure 3: *The effect of the number of views m on multiview correlation ρ and classification performance for command-ID*

as views to represent speakers. In order to test that the generalizability of speaker-ID for unseen words, we leave out 15 of the 30 speech commands during training and development (dev) and test on the left-out words. All related code is made publicly available at github.com/usc-sail/gen-dmcca.

The characteristics of the dataset are shown in Table 1. The number of views is 1123 and 15 and the number of classes is 30 and 1148 for command and speaker-ID tasks respectively. This allows us to test both conditions with a large number of views, as well as distinct classes in the signal. In all experiments, the test set does not include subjects seen in either train or dev sets.

Acoustic features: We used 30 dimensional MFCC features (extracted using Kaldi [22] with default parameters) from audio sampled at 16kHz. The frame length and frame shift were 25ms and 15ms respectively. For dMCCA, since we use CNN architecture, we also experimented with log-mel features obtained with the same configuration.

3.1. Baseline Experiments

Although the objective of our work is to analyze speech in a multiview paradigm, we compare with other methods to identify its strengths. The i-vector and x-vector methods have been extensively used for language [14] and speaker-ID [19], and as such used here as baselines. Both are trained on the Speech Commands Dataset for fair comparison with our approach. One caveat is that these methods are typically trained with vast amounts of data, which is reflected in our performance evaluation with these features.

i-vector: We separately obtain i-vectors for the command and speaker-ID tasks in the data partitions shown in Table 1. A GMM with 2048 Gaussians and a 400-dim i-vector extractor was trained. In applications with i-vectors, it is typical to apply LDA [11] to remove effects such as channel variability. Following this, for speaker-ID, the i-vectors were transformed using LDA using the ground-truth speaker identities in the training set. Similarly for command-ID, the words were used as labels. Thus, we obtain 150-dim features after LDA.

x-vector: We apply Kaldi's [22] standard x-vector recipe separately for speaker and command modeling tasks. LDA is applied to reduce the 512-dim x-vectors to 150-dim, following the standard conventions of [19].

Deep CCA: Because MCCA, and its deep version are extensions of the original CCA [2] objective, we use deep CCA [3], but trained with stochastic sampling of the views (See Eq. 3). The network architecture is the same for both deep CCA and dMCCA, shown in Fig. 2 but with only two view-branches.

Table 2: *Performance evaluation with clustering and classification; No. of classes for the two tasks are 30 and 146 respectively*

Method / performance	**command-ID**		**speaker-ID**	
	purity	macro F1	purity	macro F1
i-vector	0.52	0.69	0.33	0.44
x-vector	0.85	0.89	0.48	0.65
dCCA – MFCC	0.73	0.82	0.50	0.71
dMCCA – MFCC	0.87	0.92	0.90	0.86
dMCCA – log-mel	**0.89**	**0.94**	**0.92**	**0.90**

3.2. Generalized dMCCA

As described in Sec. 2, the different views of the data are transformed using DNNs with identical architectures, referred to as view-branches. The number of view-branches is the same as the number of subsampled views m. Although, we can use any DNN architecture, we chose CNNs because they have been successfully applied for tasks such as audio event classification [23] and speech activity detection [24]. The CNN of the view-branches in our model is a smaller version of that in [24], and is shown in Fig. 2. The input to the network with m view-branches is a $m \times T \times D$ tensor (either MFCC or log-mel) with T frames and D filter banks.

Similar to the application of temporal pooling in x-vectors, we perform global average pooling (GAP: average across outputs from convolutional filters and time-points) before input to the fully connected (FC) layer. GAP allows us to model utterances of variable duration, because the activations in time are averaged out. The activations from the FC layers, $\mathbf{H}_l$ are used to estimate the between and within-view covariances. They are additionally constrained to have a unit l_2 norm. We experimented with sigmoid and linear activations for the convolutional and FC layers and found that sigmoid activation achieved higher ρ at convergence. The number of nodes in FC layer in each view-branch was set to 64. Thus, at inference time, we obtain 64-dim features.

In all our experiments, we minimize the negative of the objective ρ using mini-batch stochastic gradient descent with a learning rate of 0.01. Additional experiments with Nesterov momentum of 0.9 and a decay of $1e-6$ improved the speed of convergence. To determine model convergence, we applied early stopping criteria (stop training if validation loss does not decrease by 1e-3 for 5 consecutive epochs).

Choosing m: A central premise of our paper is that we can generalize dMCCA to thousands of views by subsampling the views. In order to empirically study the effect of number of views m, we created ten data partitions similar to that shown in Table 1. For each one, we obtained command-ID and speaker-ID representations by varying $m = \{2, 3, 4, 5, 6, 8\}$. We also used this procedure to choose m by examining the multiview correlation ρ of the model at convergence on the dev set. The number of trainable parameters corresponding to choices of m were $\{590\text{K}, 885\text{K}, 1.1\text{M}, 1.4\text{M}, 1.7\text{M}, 2.3\text{M}\}$ which were also a factor in choosing the number of views to subsample.

3.3. Performance Evaluation

We performed classification and clustering to assess the importance of multiview representations for downstream tasks, as well as compare with the baselines. We trained independent linear support vector machines (SVM) for the representations obtained in Sec. 3.1 and dMCCA for command and speaker-

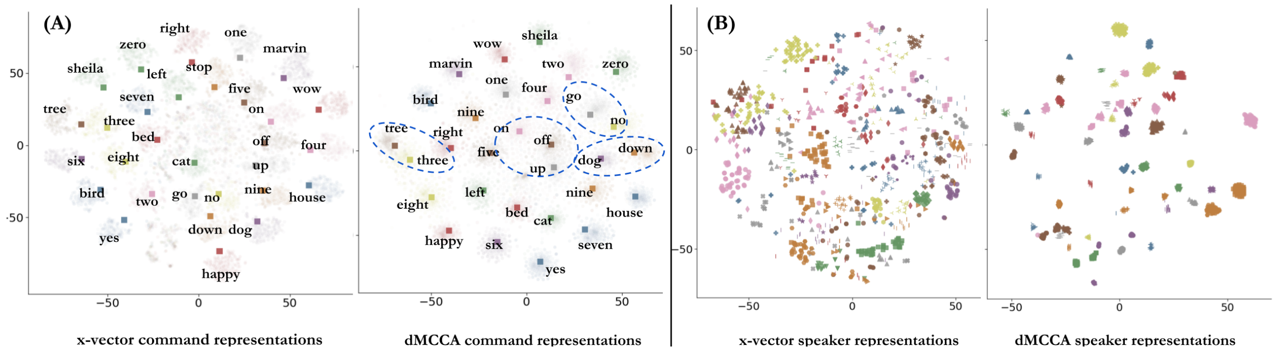

Figure 4: *t-SNE visualization of x-vectors and generalized dMCCA on the test set **(A)** command-representations: centroids are indicated by solid squares and the points belonging to a class are shaded. Notice the proximity of similar sounding words (circled in dotted lines) **(B)** speaker representations: markers+colors represent different speakers. 54 speakers with the most utterances in test-set are shown*

ID tasks. Since the representations were learned in the training set, the SVMs were trained on the dev set with 10-fold cross validation for parameter selection. Because the classes were imbalanced, we used macro-averaged F1 score to compare the models.

For unsupervised clustering, we used the implementation in sklearn [25] with kmeans++ [26] to initialize the cluster centers. The models are trained on the dev. set and the number of clusters was set to 30 and 146 for command and speaker-ID respectively during testing. We use cluster purity and V-measure [27] to assess the performance of clustering.

4. Results and Discussion

The effect of the number of subsampled views m on the multi-view correlation ρ and classification performance of these models for command-ID task is shown in Fig. 3. We observed an improvement in ρ from 0.06 to 0.25 with varying m from 2 to 3. Further increasing m only improved ρ by at most 0.02. Considering this and the relatively fewer number of trainable parameters, we chose $m = 3$ for all our subsequent experiments. The results were similar for the speaker-ID task: $\rho = 0.56 \pm 0.01$ for $m \geq 3$. Recall that ρ is bounded above by 1.

We also analyzed the effect of m on the classification performance using pairwise McNemar's [28] chi-squared test[2] with Bonferroni correction ($n = 15$ comparisons). We observed significantly better performance for $m \geq 3$ compared to $m = 2$. However the changes in performance with $m \geq 3$ were not significant. These results provide empirical evidence for stochastic sampling of views to generalize dMCCA. Our future work will focus on obtaining theoretical bounds for this result using recent analysis of streaming generalized eigenvalue problem [29].

We used t-SNE [30] to visualize the embeddings learned for commands and speakers from x-vectors as well as dMCCA. As shown in Fig. 4A, the multiview representations in the commands manifold appear invariant to the speaker characteristics. Interestingly, similar 'sounding' words are proximally located on the manifold. A few examples are {"down","dog"} and {"on", "off","up"}. The t-SNE manifold for speakers is visualized in Fig. 4B. Qualitatively, the representations from dMCCA appear to be highly separable for both ID tasks.

Next, in order to quantitatively assess if the command and speaker-ID can be performed without supervision, we compare the average purity of the clusters from kmeans approach as shown in Table 2. Similar trend in the V-measure was observed. Overall, for both tasks dMCCA performs the best. Using log-mel features instead of MFCC improved the performance. This is consistent with the use of CNNs, reported in other works [31]. While deep CCA trained with stochastically sampled views performs comparable to other baselines, dMCCA performs better suggesting the benefit of the MCCA formulation.

For command-ID, the x-vector performance is comparable to that of dMCCA, but i-vectors do not cluster as well. This is perhaps because the utterances were of small duration (about $1s$). This is consistent with findings in speaker recognition literature where i-vector system degrades as utterance length decreases [32, 33]. We observed similar results with i-vectors for speaker-ID task as well. However, compared to command-ID, x-vectors perform poorly for the speaker-ID task. This could be because the large variance in the number of utterances per speaker (See Table 1), introducing class imbalance during the supervised training of x-vectors. Similar trends in the performance were observed for classification (See Table 2).

Finally, as described in Section 2, because we obtain view-agnostic features at test time using only one of the view-branches, we evaluated the performance for the features from the other view-branches. On average we noticed a variation of 0.24 in cluster purity and 0.11 in macro-F1 which was not substantial.

5. Conclusions and Future Work

In this paper, we study speech representations in a multiview paradigm by treating the different modes of variability as multiple views using CNN and dMCCA. In the Speech Commands Dataset, we conduct two distinct experiments to identify speakers and commands where we constrain the views to the different commands or speakers respectively. We show that stochastically sampling a small number of views generalizes dMCCA for thousands of views to obtain the common information across the views, given a signal. Our performance evaluation, and comparison with other methods used to obtain robust speech representations demonstrate the benefit of explicitly modeling the different modes of variability using multiview learning. Our future work will focus on studying our approach with different *sources* of variability as views for signal processing in general.

[2] Reject H_0: marginal probabilities of each outcome are the same at $p < 0.01$

6. References

[1] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview," *Inf. Fusion*, vol. 38, no. C, pp. 43–54, Nov. 2017.

[2] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[3] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International Conference on Machine Learning*, 2013, pp. 1247–1255.

[4] K. Somandepalli, N. Kumar, R. Travadi, and S. Narayanan, "Multimodal representation learning using deep multiset canonical correlation," 2019.

[5] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*. Springer, 1992, pp. 162–190.

[6] L. C. Parra, S. Haufe, and J. P. Dmochowski, "Correlated components analysis: Extracting reliable dimensions in multivariate data," *stat*, vol. 1050, p. 26, 2018.

[7] K. Livescu and M. Stoehr, "Multi-view learning of acoustic features for speaker recognition," in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2009, pp. 82–86.

[8] S. Bharadwaj, R. Arora, K. Livescu, and M. Hasegawa-Johnson, "Multiview acoustic feature learning using articulatory measurements," in *Intl. Workshop on Stat. Machine Learning for Speech Recognition*. Citeseer, 2012.

[9] W. Wang, R. Arora, K. Livescu, and J. A. Bilmes, "Unsupervised learning of acoustic features via deep canonical correlation analysis," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4590–4594.

[10] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *CoRR*, vol. abs/1804.03209, 2018.

[11] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[12] N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, V. Hubeika, and F. Castaldo, "Support vector machines and joint factor analysis for speaker verification," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4237–4240.

[13] R. Travadi, M. V. Segbroeck, and S. S. Narayanan, "Modified-prior i-vector estimation for language identification of short duration utterances," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[14] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Twelfth annual conference of the international speech communication association*, 2011.

[15] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[16] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5115–5119.

[17] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 165–170.

[18] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, 2017, pp. 999–1003.

[19] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[20] L. C. Parra, "Multi-set canonical correlation analysis simply explained," *arXiv preprint arXiv:1802.03759*, 2018.

[21] J. J. Bartko, "The intraclass correlation coefficient as a measure of reliability," *Psychological reports*, vol. 19, no. 1, pp. 3–11, 1966.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[23] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 131–135.

[24] R. Hebbar, K. Somandepalli, and S. Narayanan, "Robust speech activity detection in movie audio: Data resources and experimental evaluation," in *Proceedings of ICASSP*, May 2019.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] D. Arthur and S. Vassilvitskii, "K-means++: the advantages of careful seeding," in *In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007.

[27] A. Rosenberg and J. Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning(EMNLP-CoNLL)*, 2007, pp. 410–420.

[28] W. D. Dupont and W. D. Plummer, "Power and sample size calculations: a review and computer program," *Controlled clinical trials*, vol. 11, no. 2, pp. 116–128, 1990.

[29] K. Bhatia, A. Pacchiano, N. Flammarion, P. L. Bartlett, and M. I. Jordan, "Gen-oja: A simple and efficient algorithm for streaming generalized eigenvector computation," *CoRR*, vol. abs/1811.08393, 2018.

[30] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," 2008.

[31] C.-W. Huang and S. Narayanan, "Characterizing types of convolution in deep convolutional recurrent neural networks for robust speech emotion recognition," *CoRR*, vol. abs/1706.02901, 2017.

[32] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, and M. W. Mason, "I-vector based speaker recognition on short utterances," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association*. International Speech Communication Association (ISCA), 2011, pp. 2341–2344.

[33] A. Jati and P. Georgiou, "Neural predictive coding using convolutional neural networks towards unsupervised learning of speaker characteristics," *Accepted in IEEE Transactions on Audio, Speech and Language Processing, arXiv preprint arXiv:1802.07860*, 2018.