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

Unsupervised low-rank representations for speech emotion recognition

Georgios Paraskevopoulos[1,2], *Efthymios Tzinis*[3], *Nikolaos Ellinas*[1], *Theodoros Giannakopoulos*[2], *Alexandros Potamianos*[1,2]

[1]School of Electrical & Computer Engineering, National Technical University of Athens, Greece
[2]Behavioral Signal Technologies, Los Angeles, CA, USA
[3]Department of Computer Science, University of Illinois at Urbana-Champaign, IL, US

geopar@central.ntua.gr, etzinis@gmail.com, nellinas@central.ntua.gr, thodoris@behavioralsignals.com, potam@central.ntua.gr

Abstract

We examine the use of linear and non-linear dimensionality reduction algorithms for extracting low-rank feature representations for speech emotion recognition. Two feature sets are used, one based on low-level descriptors and their aggregations (IS10) and one modeling recurrence dynamics of speech (RQA), as well as their fusion. We report speech emotion recognition (SER) results for learned representations on two databases using different classification methods. Classification with low-dimensional representations yields performance improvement in a variety of settings. This indicates that dimensionality reduction is an effective way to combat the curse of dimensionality for SER. Visualization of features in two dimensions provides insight into disriminatory abilities of reduced feature sets.

Index Terms: Non-linear, dimensionality reduction, emotion recognition, speech, manifold learning, autoencoder

1. Introduction

Human-machine interaction is constantly evolving towards the use of more natural interfaces, like speech. Still the key difference between human-human and human-machine communication is the ability of humans to recognize the emotion of their conversation peers and modify their communication strategy based on that. Although significant progress has been done in the field of speech emotion recognition (SER), machines have not achieved human-like performance. One of the reasons is the scarcity of available annotated data. SER databases are mostly composed of relatively small number of utterances from few speakers, which limits the generalization abilities of the models. Furthermore modern SER systems rely on feature sets of high dimensions. The small amount of training samples do not cover all combinations of values in the high-dimensional feature spaces and, thus, SER algorithms suffer from the curse of dimensionality (CoD) [1]. In this work we postulate that reducing dimensionality of the feature space is an effective way to combat CoD and demonstrate that low-dimensional representations yield simpler models with comparable performance. Dimensionality reduction (DR) algorithms aim at learning low-dimensional latent representations of real world data. Such representations can be used for exploratory data analysis, to visualize and gain intuition on the statistical properties of data or, as in our case, extract latent features for input to classification or regression models.

Evidence that DR on speech features can create robust representations for SER can be found in the literature. In [2], Principal Component Analysis (PCA) [3] is used to extract low-dimensional representations for the feature set introduced in [4] containing 6552 features. The system is evaluated on Berlin emotional database (Emo-DB) [5]. In [6] Linear Discriminant Analysis (LDA) [7] and PCA are used for SER, along with a weighted variation of LDA on a feature set of 225 dimensions. Experiments showed no significant performance difference between PCA and LDA. These methods are also compared in [8] and [9], along with Sequential Forward Selection (SFS) [10], on a feature set consisting of 48 prosodic features and 16 formants. PCA representations extracted in [8] are found to be inferior than LDA, while [9] observed no significant difference. SFS and PCA are also explored in [11] for the Danish Emotion Speech database [12]. [13] experimentally found that applying PCA on utterance-level statistics of pitch and energy features gives equivalent SER performance with the original features on a call center dialog corpus. Authors in [14] report that classification accuracy keeps improving when increasing the number of principal components only up to a centain rank for a feature set of 33 dimensions. A supervised variation of PCA along with Greedy Feature Selection (GFS) [15] and Elastic-Net [16] are explored in [17] on two sets of energy-based and MFCC-based feature sets of 400 and 82 dimensions, with inconclusive results as to which approach is superior. The application of Linear and non-linear DR methods on SER is examined on a prosodic feature set of 48 dimensions in [18]. Compared methods include unsupervised methods like PCA, Isometric Mapping (ISOMAP) [19] and Locally Linear Embedding (LLE) [20], and supervised methods LDA, Supervised LLE (SLLE) [21], Neighborhood Component Analysis (NCA) [22], Maximally Component Metric Learning (MCML) [23], local Fisher Discriminant Analysis (LFDA) [24] and Modified SLLE (MSLLE). Results show better performance of PCA for unsupervised DR while MSLLE was superior for supervised DR.

2. Dimensionality Reduction Algorithms

DR algorithms compress data in a low-dimensional space while preserving meaningful statistical and geometrical properties. Such properties are covariance of original data, pairwise distances between samples or local neighborhoods. They can be separated into two general categories, linear and non-linear. Linear DR aims to find a linear projection $Y = TX \in R^{n \times k}$ of the real data $X \in R^{n \times m}$, where $k < m$. Examples of linear DR algorithms are PCA and classical multidimensional scaling (cMDS) [25]. PCA projects data into a low-dimensional space, which is formed by an orthogonal basis of linearly uncorrelated vectors called the principal components. Principal components are selected as the axes along which the samples have maximum variance. cMDS takes a geometric approach, finding a set of low-dimensional points that best preserve pairwise euclidean distances between original data points.

http://dx.doi.org/10.21437/Interspeech.2019-2769

Non-linear dimensionality reduction (NLDR) algorithms aim to infer the intrinsic geometry of the original data, based on the manifold hypothesis, which states that real world data tend to lie on a low-dimensional manifold, embedded in the high-dimensional space. These algorithms are not limited in linear transformations, like the rotations and stretches that can be induced by a matrix multiplication. An extension of cMDS is metric MDS [26] where dissimilarity measures are assumed metric, but not necessarily euclidean. When these measures are closely related to the euclidean distance, e.g. cosine distance, metric MDS is still characterized as a linear DR approach. Stress Majorization [26] and Pattern Search MDS [27] are two algorithms for metric MDS. The non-metric extension of MDS [28] tries to approximate the rank order of original distances by applying a monotonically increasing function, usually approximated by isotonic regression. ISOMAP finds an isometric mapping of the original data by extending metric MDS to approximate geodesic pairwise distances between original samples space as euclidean pairwise distances in the transformed samples. Geodesic distances are approximated by the shortest path distances between data points. While MDS and ISOMAP consider the global data geometry, Local Linear Embedding (LLE) reconstructs local regions by finding sets of weights which are used to represent samples as a weighted combination of their closest neighbors. Representations are computed by solving a sparse eigenvalue problem. Modified LLE [29] is an extension of LLE that uses multiple neighborhood weights and produces more robust results. Another non-linear approach is Laplacian Eigenmaps or Spectral Embedding [30], which preserves local manifold geometry by minimizing the Laplacian of the graph formed by neighboring data points. The Laplacian of this graph approximates the Laplacian-Beltrami operator over the manifold, which indicates the divergence of the mapping of a high-dimensional point to the low-dimensionsional manifold.

Autoencoders [31] are a class of deep neural networks that can be used for linear and non-linear dimensionality reduction and are composed of an encoder and a decoder. Encoder projects input x to a low-dimensional space via a hidden layer h, while the attempts to reconstruct x from h. If no non-linear activations are used, encoder learns a linear projection $Wx + b$, whereas if we use a non-linear activation function (e.g. sigmoid or rectified linear unit) in the output of the encoder's layers, a non-linear embedding is learned.

3. Features for speech emotion recognition

We consider the following feature sets:

IS10 set: The IS10 feature set [32] consists of 1582 features. IS10 is obtained by transforming the signal in the Fourier space. Features correspond to 21 statistical functionals (e.g. percentiles, linear regression coefficients) applied to 38 low level descriptors (MFCCs, PCM loudness etc.) and their deltas. Extraction is performed using the openSMILE toolkit.

RQA set: The Recurrence Quantification Analysis (RQA) feature set [33] consists of 432 features. This feature set is obtained by analyzing speech dynamics through phase space representation. The phase space is reconstructed through the use of time-delayed versions of the original signal and then the recurrence plots are calculated as thresholded pairwise distances of points in the phase space. Features are extracted as aggregated RQA measures from the recurrence plots. Source code for feature extraction is publicly available.[1]

Fused set: We concatenate features from IS10 and RQA into a represenation of 2014 dimensions, modeling both frequency content of speech signals and recurrence dynamics.

4. Experiments and Results

4.1. Experimental Setup

We use the following databases for evaluation:

Emo-DB: Berlin Database of Emotional Speech (Emo-DB) [5] contains 535 emotional German sentences, voiced by 10 actors (5 male and 5 female). Specifically, 7 emotions are included i.e., 127 anger, 45 disgust, 70 fear, 71 joy, 60 sadness, 81 boredom and 70 neutral.

IEMOCAP: IEMOCAP database [34] contains 12 hours of video data with scripted and improvised dialog recorded by 10 actors. Utterances are organized in 5 sessions of dyadic interactions between pairs of actors. For our experiments we consider 5531 utterances of 4 emotions (1103 angry, 1636 happy, 1708 neutral and 1084 sad), where we merge excitement class into happiness [35], [36], [37], [38].

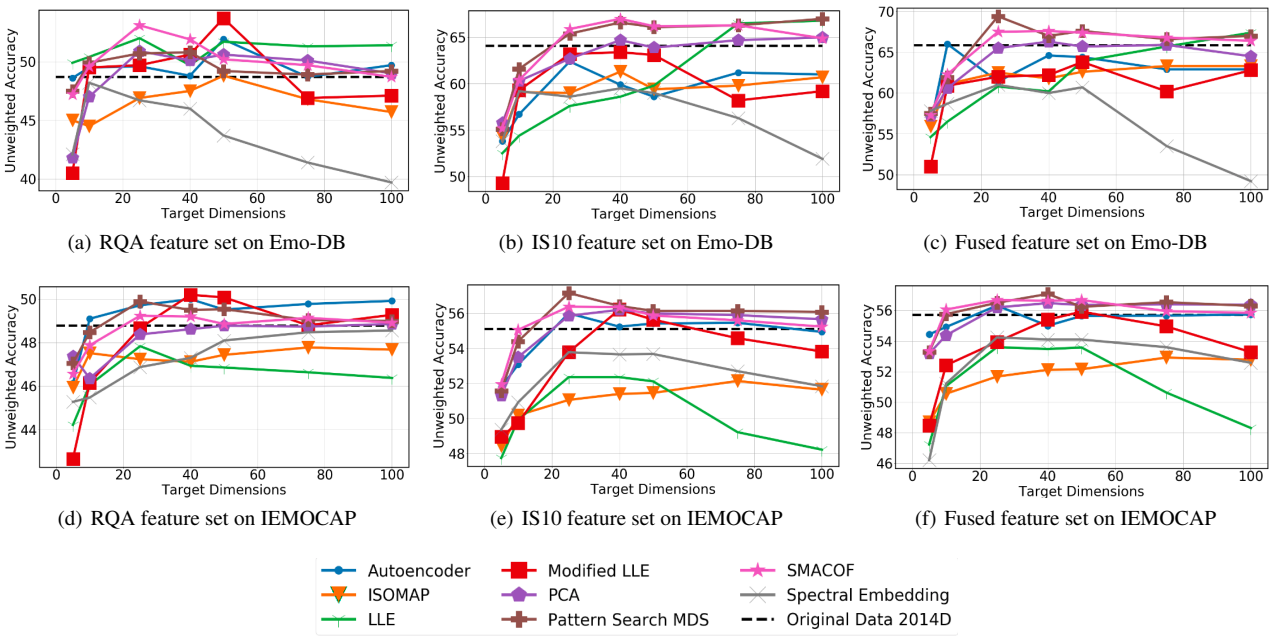
We consider utterance-level, speaker independent (SI) SER for our experiments. In this setup a number of speakers are kept hidden from the training set and used for evaluation. Specifically in the case of Emo-DB we perform leave one speaker out (LOSpO) cross-validation, where test folds contain the instances of the unknown speaker. For IEMOCAP we use the leave one session out (LOSO) cross-validation scheme, where two speakers participating in a session are used as the evaluation folds. This results in a 10-fold cross-validation scheme for Emo-DB and 5-fold cross-validation for IEMOCAP. We apply Z-normalization to standardize the features in zero mean and unit variance, where each sample x is transformed according to the formula $z = \frac{x-\mu}{\sigma}$. Note that for SI experiments only samples in the training set are used to calculate μ and σ and test samples are normalized using these statistics.

Representations resulting from all DR approaches are evaluated for k-nearest neighbors (kNN) classification. We perform grid search on the optimal number of neighbors k in the $[1, 30]$ range and report results for the optimal value for each dimension and each method. Optimal values of k range from 13 to 20 indicating that consistent neighborhoods are formed in the low-dimensional spaces. We also evaluate low-rank representations on SVM with linear and gaussian kernels, and Logistic Regression (LR), with optimal value of C in the range $[0.01, 10]$. Autoencoder is trained with 3 encoder layers, 3 decoder layers and 1 hidden layer, using ReLU activations.

4.2. Results

As evaluation metrics we used both weighted accuracy and unweighted accuracy. For brevity we report unweighted accuracy results, noting that same trends form with respect to the weighted accuracy metric. Fig. 1(a) shows the results of DR applied to the RQA features on Emo-DB for all DR methods, for different embedding dimensions L. We observe that Modified LLE achieves the best results when $L = 50$, followed by SMACOF MDS in $L = 25$. Observe that all methods except ISOMAP and Spectral Embedding manage to outperform the original features of 432 dimensions. In Fig. 1(b), which shows results for DR on IS10 features for Emo-DB, we can

[1] https://github.com/etzinis/nldrp



(a) RQA feature set on Emo-DB (b) IS10 feature set on Emo-DB (c) Fused feature set on Emo-DB

(d) RQA feature set on IEMOCAP (e) IS10 feature set on IEMOCAP (f) Fused feature set on IEMOCAP

Figure 1: *Results of DR for different feature sets on IEMOCAP and Emo-DB*

observe a different pattern. Here the MDS algorithms perform best for every embedding dimension, followed by PCA, all three of these methods outperforming the original feature set of 1582 dimensions. This indicates that this feature set resembles more a hyperplane in the high-dimensional space than a non-linear manifold. Non-linear methods like LLE, ISOMAP and Spectral Embedding underperform. For the fused feature set in Fig. 1(c) we see again that distance-preserving transformations yield the best performance. Same patterns emerge in IEMOCAP in Fig. 1(d), 1(e), 1(f), with Modified LLE achieving better performance for the RQA features and Pattern Search MDS and PCA yielding best representations for IS10 features. Notably in IEMOCAP, performance of the Autoencoder is significantly better because there are more training samples. For the experiments with the fused feature set we again observe a consistent pattern in both Emo-DB and IEMOCAP, with MDS yielding again the best representations followed by PCA. Fusion is still beneficial after applying DR though we observe that the structure of IS10 features dominates under fusion.

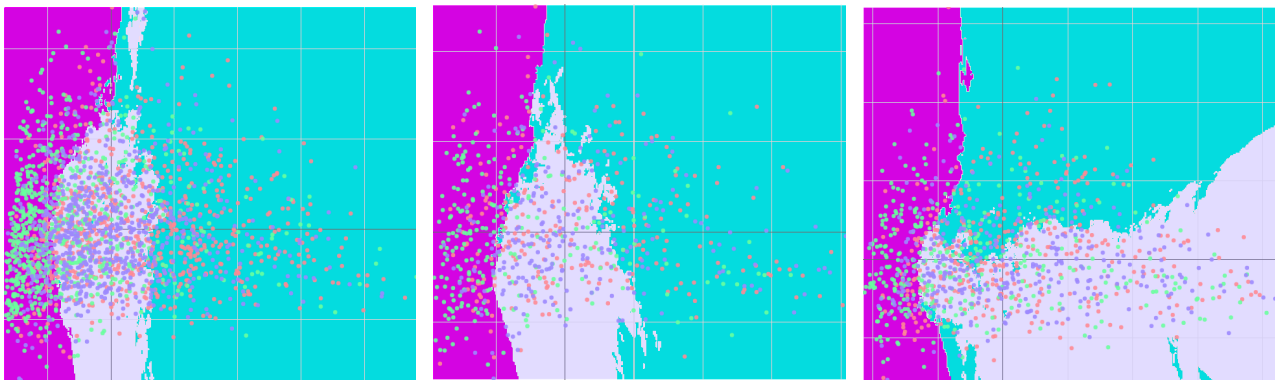
In Table 1 we show results for linear SVM, radial basis function (rbf) SVM, kNN and LR. We reduce dimensionality of IS10 features from 1582 to 25 dimensions and report unweighted accuracy (UA) on IEMOCAP. Low-rank representations produce very competitive results to the original sparse features, while for linear SVM and kNN they even improve classification accuracy. Overall global, linear DR methods like MDS and PCA produce the best representations.

4.3. Visualization

We include visualizations of feature maps reduced in 2D. We focus on the best and the worst performing methods and comment on some interesting observations.

Figure 2 demonstrates the results of PCA into two dimensions, for a large proprietary and internally annotated dataset containing speech segments from multiple domains such as movies, TV series and interviews. Subfigures illustrate the distributions of the speech segments into the two PCA dimensions for three emotional classes: anger, happiness and sadness. In addition, we illustrate the decision surfaces for a simple kNN classifier. The results demonstrate how the blue class (anger) is similarly distributed between the red and green (sadness and happiness respectively) for the two first domains (Series and Movies) in Fig 2(a) and Fig. 2(b) respectively, based on the primary PCA dimension (x axis). On the other hand, for the interviews domain, the primary PCA dimension is not enough to discriminate between the emotional classes as we see in Fig. 2(c). On the contrary, the anger and happiness classes are mostly discriminated based in the second PCA dimension. Interestingly, this unsupervised distribution is quite similar to the Valence-Arousal affective representation. This example demonstrates how an unsupervised dimensionality reduction can be very sensitive to changes in domain when illustrating emotional content.

Fig. 3(a) shows the 2D space created using Pattern Search MDS, which maps the points inside an elongated disk area. We can see on the left the anger points while the sadness points are on the right. Close to anger is happiness samples, while boredom is close to sadness. Other emotions lie in the middle. So it looks like that even in the 2D space MDS learns meaningful representations, with x axis being a latent feature that can encode arousal. On the contrary LLE, which tries to preserve local neighborhoods and yields poor recognition accuracy on the fused feature set concentrates most samples in the center as we can see in Fig. 3(b), but still we can observe low arousal emotions (sadness) being separated from high arousal ones (anger). In Fig. 4 we show ISOMAP embeddings for two speakers in IEMOCAP. Observe, although ISOMAP cannot separate emotions, it achieves a better discrimination result, in terms of speaker separation, for this experiment. One could consider basing a speaker diarizer on geodesic distances between samples.

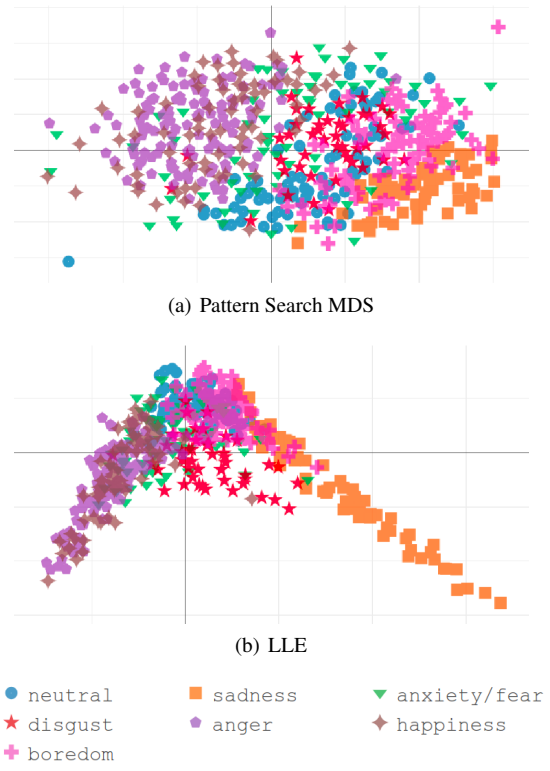


(a) Series (b) Movies (c) Interviews

Figure 2: *Cross-domain decision regions with 2D DR*

Table 1: *Classification on IS10 features for IEMOCAP (UA)*

	SVM (linear)	SVM (rbf)	kNN	LR
Pattern S. MDS	**56.0**	57.5	56.5	55.4
SMACOF MDS	55.8	**58.5**	**56.7**	**55.8**
PCA	55.8	57.7	56.2	**55.8**
ISOMAP	52.3	52.5	51.7	52.2
LLE	53.4	54.2	53.6	53.2
Modified LLE	54.6	47.0	53.9	55.5
Spectral Emb.	54.1	54.3	54.2	55.1
Autoencoder	55.4	57.8	56.3	55.5
Original 1582D	54.7	**59.8**	55.7	**56.9**



(a) Pattern Search MDS

(b) LLE

Figure 3: *2D DR for fused feature set on Emo-DB*

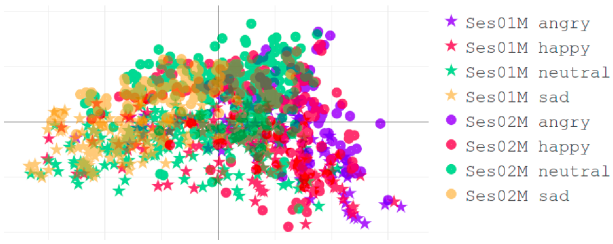


Figure 4: *Isomap on fused features for 2 IEMOCAP speakers*

5. Conclusions

In this work we explore the effects of unsupervised linear and non-linear DR on state-of-the-art speech features for SER. We evaluate these algorithms for speaker independent SER on IEMOCAP and Emo-DB. Experiments show that performance of low-rank representations is competitive to original high-dimensional representations. This phenomenon is hypothesized to be caused by the curse of dimensionality, since the number of samples in SER datasets does not span the high-dimensional space. Interpretation of results and vizualization of 2D representations gives interesting insights on the high-dimensional structures. First insight is that IS10 features can be decomposed by use of linear DR, e.g. by use of PCA or MDS algorithms. Second, distance preserving DR can encode meaningful dimensions, e.g. arousal. Third, speaker samples can be separated by isometric mappings. Fourth, unsupervised DR can be rather sensitive when illustrating cross-domain emotional content. Future work will focus on creating end-to-end representations using autoencoders with distance preserving regularization and investigating the interesting insight on using geodesic-distance preserving representations for speaker separation.

6. Acknowledgements

This work has been partially supported by computational time granted from the Greek Research & Technology Network (GR-NET) in the National HPC facility - ARIS and the EU-IST H2020 BabyRobot project under grant #687831.

7. References

[1] R. Bellman, *Adaptive control processes: a guided tour*. Princeton University Press, 2015, vol. 2045.

[2] B.-C. Chiou and C.-P. Chen, "Feature space dimension reduction in speech emotion recognition using support vector machine," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2013, pp. 1–6.

[3] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[4] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Proc. IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2009, pp. 552–557.

[5] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. INTERSPEECH*, 2005, pp. 1517–1520.

[6] J. Yuan, L. Chen, T. Fan, and J. Jia, "Dimension reduction of speech emotion feature based on weighted linear discriminate analysis," *Image Processing and Pattern Recognition*, vol. 8, pp. 299–308, 2015.

[7] R. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[8] M. You, C. Chen, J. Bu, J. Liu, and J. Tao, "Emotion recognition from noisy speech," in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, 2006, pp. 1653–1656.

[9] ——, "A hierarchical framework for speech emotion recognition," in *Proc. IEEE International Symposium on Industrial Electronics*, 2006, pp. 515–519.

[10] K. Fu, *Sequential methods in pattern recognition and machine learning*. Academic Press, 1968, vol. 52.

[11] D. Ververidis, C. Kotropoulos, and I. Pitas, "Automatic emotional speech classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2004, pp. 593–596.

[12] I. Engberg, A. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a danish emotional speech database," in *Proc. Fifth European Conference on Speech Communication and Technology (EUROSPEECH)*, 1997, pp. 1695–1698.

[13] C. Lee, S. Narayanan, and R. Pieraccini, "Classifying emotions in human-machine spoken dialogs," in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, 2002, pp. 737–740.

[14] Z.-J. Chuang and C.-H. Wu, "Emotion recognition using acoustic features and textual content," in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, 2004, pp. 53–56.

[15] A. Farahat, A. Ghodsi, and M. Kamel, "An efficient greedy method for unsupervised feature selection," in *Proc. IEEE 11th International Conference on Data Mining (ICDM)*, 2011, pp. 161–170.

[16] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, p. 1, 2010.

[17] P. Fewzee and F. Karray, "Dimensionality reduction for emotional speech recognition," in *Proc. ASE/IEEE International Conference on Privacy, Security, Risk and Trust (PASSAT) and International Confernece on Social Computing (SocialCom)*, 2012, pp. 532–537.

[18] S. Zhang and X. Zhao, "Dimensionality reduction-based spoken emotion recognition," *Multimedia Tools and Applications*, vol. 63, no. 3, pp. 615–646, 2013.

[19] J. Tenenbaum, V. d. Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[20] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[21] D. De Ridder, O. Kouropteva, O. Okun, M. Pietikäinen, and R. Duin, "Supervised locally linear embedding," in *Artificial Neural Networks and Neural Information Processing ICANN/ICONIP*, 2003, pp. 333–341.

[22] J. Goldberger, G. Hinton, S. Roweis, and R. Salakhutdinov, "Neighbourhood components analysis," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2005, pp. 513–520.

[23] A. Globerson and S. Roweis, "Metric learning by collapsing classes," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2006, pp. 451–458.

[24] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *Journal of Machine Learning Research*, vol. 8, no. May, pp. 1027–1061, 2007.

[25] W. Torgerson, "Multidimensional scaling: I. theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952.

[26] J. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[27] G. Paraskevopoulos, E. Tzinis, V. E., and P. A., "Pattern Search Multidimensional Scaling," 2018, arXiv:1806.00416v2.

[28] J. B. Kruskal, "Nonmetric multidimensional scaling: a numerical method," *Psychometrika*, vol. 29, no. 2, pp. 115–129, 1964.

[29] Z. Zhang and J. Wang, "Mlle: Modified locally linear embedding using multiple weights," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2007, pp. 1593–1600.

[30] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2002, pp. 585–591.

[31] D. Ballard, "Modular learning in neural networks," in *Proc. Sixth National Conference on Artificial Intelligence*, 1987, pp. 279–284.

[32] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. INTERSPEECH*, 2010, pp. 2794–2797.

[33] E. Tzinis, G. Paraskevopoulos, C. Baziotis, and A. Potamianos, "Integrating recurrence dynamics for speech emotion recognition," *Proc. INTERSPEECH*, pp. 927–931, 2018.

[34] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, p. 335, 2008.

[35] Z. Aldeneh and E. Provost, "Using regional saliency for speech emotion recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2741–2745.

[36] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Transactions on Affective Computing*, no. 1, pp. 3–14, 2017.

[37] H. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[38] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition," in *Proc. INTERSPEECH*, 2016, pp. 3603–3607.