Acta Acustica united with Acustica
Vol. 104 (2018) 737 – 740

DOI 10.3813/AAA.919214

Emotion Recognition in Speech with Latent Discriminative Representations Learning

Jing Han[1], Zixing Zhang[2], Gil Keren[1], Björn Schuller[1,2]

[1] ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany. jing.han@informatik.uni-augsburg.de

[2] Group on Language, Audio & Music, Imperial College London, UK

Summary

Despite significant recent advances in the field of affective computing, learning meaningful representations for emotion recognition remains quite challenging. In this paper, we propose a novel feature learning approach named Latent Discriminative Representation (LDR) learning for speech emotion recognition. Unlike most existing hand-crafted features designed for specific applications or features learnt by a standard neural network, the proposed learning method incorporates an additional training objective in order to learn better representations of the task of interest. To this end, we group the training samples into sets of triplets, satisfying that the second member in each triplet comes from the same class as the first and that the third member comes from a dif ferent class than the first. In the training pr ocess, we maximise the distance of the samples from different classes in the latent representation space, while we minimise the distance for samples from the same class. To evaluate the effectiveness of LDR, we perform extensive experiments on the widely used database IEMOCAP, and find that the LDR improves performance over the standard neural network training procedure.

PACS no. 43.72.Ar

1. Introduction

Over the past few decades, massive efforts have been made to extract hand-crafted acoustic features that can capture relevant information for specific tasks. For instance, cochleagram-based features have been applied to recognise emotional vocalisations of canine [1], whereas emotional articulatory changes have been shown to be the most effective features to distinguish happiness from anger [2]. However, in most cases, appropriate and strong domain knowledge is required to design a suitable feature set for tasks at hand. To address this limitation, a large variety of methods emerged recently to use neural networks for learning more generic representations directly from the raw data, such as Convolutional Neural Networks and Recurrent Neural Networks (RNNs).

Moreover, learning meaningful representations is an important task for Speech Emotion Recognition (SER) where specific domain knowledge is rather vital. Recently, various methods have been proposed to design an appropriate representation of the speech data for promising SER performance, such as autoencoder-based methods, shared-hidden-layer-based representation learning, and deep spectrum features. While these methods have produced useful representations for SER, none have explicitly utilised training strategies to elicit semantic structure in the latent representation space. Contrary to these works, we propose taking the supervisory information in form of class labels into consideration during training, to preserve semantic structure of the data while learning the representations. To this end, we present a feature learning paradigm called Latent Discriminative Representation (LDR) learning. To preserve semantic structure of the data, we maximise the diversity of training samples belonging to various classes, while retaining the similarity of training samples within the same class. With this goal, LDR learning can be deemed as an optimisation problem of pairwise relation based loss function, where the semantic relation can be incorporated into the similarity and dissimilarity among pairs of instances. As a consequence, this could provide a latent discriminative learnt feature-space to ameliorate the classification performance. In this work, experiments are carried out to generate LDR features from hand-crafted features for SER on the IEMOCAP database, and results show that our proposed LDR features yield performance improvement over the traditional hand-crafted features.

Received 20 June 2018,
accepted 4 September 2018,
published online 12 September 2018.

2. Related Work

Recently, the Deep Structured Semantic Model (DSSM) [3] has been investigated and achieved appealing performances for many text processing tasks. In DSSM, vector representations are generated in a continuous semantic space where semantic similarity of two text strings can be modelled. Specifically, strings are projected into a low-dimensional space in which the relevance of two strings can be indicated easily as the cosine similarity between their semantic representations. Motivated by the success of DSSM, this work advocates to generate a latent discriminative feature space in which the semantic relationship can be estimated by distance in the learnt space.

Our proposed learning method is further closely related to triplet networks [4] that have been leveraged to learn useful semantic representations in computer vision tasks, by distinguishing similar and dissimilar pairs of training instances. In the related field of acoustic signal processing, however, only a few works have been reported towards this direction very recently. In [5], a triplet loss function was employed to project i-vectors into a space that better separates speakers in terms of cosine similarity, using a simple feed-forward neural network. Similar work has been done in [6]. Overall, one may notice that these methods are mainly applied to speaker embedding and verification, which differ from our framework that is particularly designed for emotion prediction.

3. Methodology

The overview of the latent discriminative representation learning scheme is shown in Figure 1. The original feature vector $\mathbf{x}$ is fed into a Deep Neural Network (DNN) with multiple hidden layers to generate the corresponding LDR $\mathbf{l}$, i. e., the output from the representation layer of the network. The procedure can be denoted by $f(\mathbf{x}) \in \mathbb{R}^K$, meaning that it embeds the vector $\mathbf{x}$ into a K-dimensional space. When training, to enforce instances from the same class to be closer in representation space as well as retaining the different classes a larger distance, we propose to use a set of three instances from the training set as an input triplet for each run of training. In what follows, we describe the LDR learning approach in detail, using the annotation illustrated in Figure 1. In the figure, the network is unfolded three times and placed in parallel for a better view and explanation.

Formally, given a set of triplets $\tau = \{\tau_i\}_{i=1}^n$, and $\tau_i = \{\mathbf{x}_i, \mathbf{x}_i^+, \mathbf{x}_i^-\}$, where $\mathbf{x}_i$ and $\mathbf{x}_i^+$ (denoted as a *positive pair*) are from the same class, and $\mathbf{x}_i$ and $\mathbf{x}_i^-$ (denoted as a *negative pair*) belonging to different classes, our target is to learn a mapping to a latent representation space where $\mathbf{x}_i$ is more similar (or closer) to $\mathbf{x}_i^+$ than to $\mathbf{x}_i^-$. processing instances in τ, the latent feature representations $\mathbf{l}_i$, $\mathbf{l}_i^+$, and $\mathbf{l}_i^-$ can be obtained from the representation layer for $\mathbf{x}_i$, $\mathbf{x}_i^+$, and $\mathbf{x}_i^-$, respectively. Based on these, the distance of the positive pair D_i^+ and the distance of the negative pair D_i^- are computed as

$$D_i^+ = \left\| \mathbf{l}_i - \mathbf{l}_i^+ \right\|_2 = \left\| f(\mathbf{x}_i) - f(\mathbf{x}_i^+) \right\|_2, \qquad (1)$$

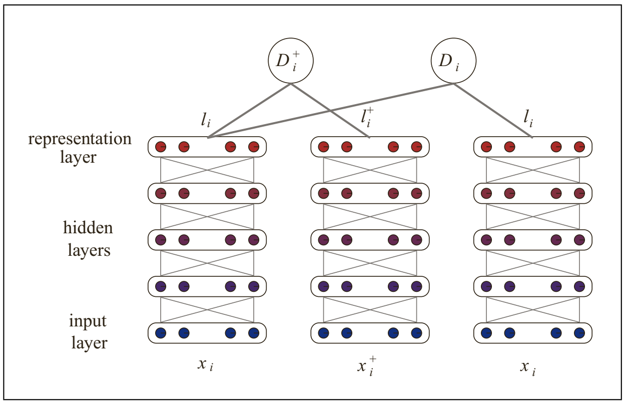

Figure 1. Framework for learning latent discriminative representations. For each training instance $\mathbf{x}_i$, $\mathbf{x}_i^+$ indicates a randomly selected instance in the same category as for $\mathbf{x}_i$; $\mathbf{x}_i^-$ indicates another randomly selected instance from a different category; $\mathbf{D}_i^+$ and $\mathbf{D}_i^-$ respectively denote the distances between the two latent representations learnt from the instances with the same or different categories.

$$D_i^- = \left\| \mathbf{l}_i - \mathbf{l}_i^- \right\|_2 = \left\| f(\mathbf{x}_i) - f(\mathbf{x}_i^-) \right\|_2, \qquad (2)$$

where $\|\cdot\|_2$ denotes the Euclidean distance between the two LDRs in a pair.

In training, the model parameters are estimated to encourage instances with the same label to approach each other and instances with different labels to be apart from each other. Equivalently, the objective of the target problem can be deemed as decreasing D_i^+ and meanwhile increasing D_i^- over all triplets. For this purpose, we try to minimise the loss function that is defined as

$$\mathcal{L}_\Lambda = b^d, \qquad (3)$$

where

$$d = D_i^+ - D_i^-, \qquad (4)$$

d indicates the discrepancy between the positive and negative pairs, b is a predefined base of the exponential function ($b > 1$ to ensure exponential growth), and Λ denotes the trainable parameters of the network. Since $\mathcal{L}_\Lambda$ is differentiable with respect to Λ, the loss function in Equation (3) can be readily integrated in back propagation in neural networks.

Here, in contrast to taking merely the discrepancy d as the objective, in Equation (3) the exponential function is introduced. The underlying rationale is that we aim to learn more useful representations, by giving different triplets various emphasis mainly depending on the difficulty of distinguishing contents of the triplet in the learnt space. In other words, instead of paying equal attention to each triplet during training, attentions are of exponential growth with respect to d. Mathematically, when there is a large discrepancy d, it results in an even larger gradient of $\mathcal{L}_\Lambda$ to update the network; if d is small, the network updates its weights only slightly. Therefore, emphasis is particularly placed to enforce the representation learning to facilitate difficult and ambiguous triplets, including cases when $\mathbf{l}_i^-$ is close to $\mathbf{l}_i$, when $\mathbf{l}_i^+$ is far from $\mathbf{l}_i$, or in cases when $\mathbf{l}_i$ is closer to $\mathbf{l}_i^-$ than t is to $\mathbf{l}_i^+$. As a result, the model learns to project the original features into a latent space where the intra-class feature distance is smaller comparing with the inter-class feature distance. Once training is completed, a new instance $\mathbf{x}_t$ can be fed into the trained network to generate its corresponding LDR feature $\mathbf{l}_t$ for further processing.

Additionally, the learning scheme can be extended to further concern relative distance differences in a multi-class scenario where we can apply tuples of $n + 1$ for n-class classification in place of triplets. That is, one positive pair and another $n - 1$ negative pair of samples can be formed for each training sample. Then, Equation (4) needs to be adjusted to take all discrepancies into account. This will lead to an embedding space that may achieve even better classification performance compared with the triplet loss. However, when the number of the total categories is large, it may result in a high computational requirement. In this work, we focus our analysis on triplets for the sake of reducing the computational complexity.

Furthermore, it is worth noting that the proposed LDR learning scheme can be applied to different DNN structures for specific tasks. In this work, we focus on SER and utilise RNNs with Long Short-Term Memory cells as they were found to yield good overall performance in SER [7].

4. Experiments

In this section, we implement and evaluate our approach for speech emotion classification.

4.1. Dataset and Features

To validate the proposed paradigm, we used the Interactive Emotional dyadic MOtion CAPture (IEMOCAP) database, which contains approximately 12 hours of recordings from five pairs of experienced actors [8]. The recordings were then segmented into utterances and further annotated both in nine categorical emotions (anger, sadness, happiness, disgust, fear, surprise, frustration, excitement and neutral states) and in two dimensional aspects, i. e., activation and valence, on a five-point scale. In this work, we divided the dataset into three speaker independent partitions, i. e., 6 319 for training, 1 811 for development, and 1 819 for test.

To extract acoustic features from the segments, the openSMILE toolkit was used to extract a minimalistic expert-knowledge based feature set, i. e., eGeMAPS [9], which contains 88 statistical features calculated by applying various functionals over 23 Low-Level Descriptors.

4.2. Implementation Details

For activation and valence, to keep in line with other works in the literature [10, 11], we projected the five-point scale for dimensional labels into three-points. For the emotional state, we only considered four categories, i. e., happiness, sadness, anger, and neutral, since all other categories appear very sparsely in the dataset. We firstly carried out the baseline experiments, where a classifier was trained on the original hand-crafted feature sets (cf. Section 4.1) for each task (i. e., activation, valence, or emotion), separately. Specifically, we used a linear Support Vector Machine (SVM) and a RNN, mainly due to their widespread usage and appealing performance achieved in emotion recognition [7]. For SVM, the complexity of the SVM was optimised on the development set via searching between .0001 and 5. Similarly, for the RNN, the structure was determined on the best performance achieved on the development set via a grid search over [1, 2, 3, 4, 5] recurrent layers and [250, 500, 1000, 2000] hidden units per layer. When training neural networks, the weights were updated for every minibatch of 64 instances, and performance was evaluated on the development set for every 50 iterations within a maximum of 10000 iterations, that were enough to reach convergence.

To learn the LDR model, we fed the original features into another RNN, which was trained by the proposed training strategy as described in Section 3. For the sake of simplicity, we kept each hidden layer and the representation layer with the same number of hidden units. Again, the structure of this network was optimised on the development set by a grid search over [1, 2, 3, 4, 5] for the number of recurrent layers and [250, 500, 1000, 2000] for the number of hidden units per layer. Additionally, the RNN was trained with an Adam optimiser with an initial learning rate of 10^{-4}.

Finally, to measure the performance of the systems, we utilised the frequently used metrics of F1 and Unweighted Average Recall (UAR – the sum of classwise recall divided by the number of classes) for SER.

4.3. Results and Discussion

For our experiments, we conducted three prediction tasks, i. e., activation, valence, and emotion classifications on audio signals. In all these experimental scenarios, Table I presents the performance of the models using the original features or the learnt LDRs, on both the development and test set. From the table, we can observe that the learnt LDRs outperform the traditional hand-crafted features in most of the scenarios measured by F1 or UAR. More specifically, for activation prediction, compared with the original hand-crafted features, LDRs yield higher F1 and UAR on both the development and test sets. Similar observations can be made for both valence and emotion predictions.

Additionally, comparing the performance achieved by LDR-SVM and LDR-RNN on the test set, it is noticed that, LDR-RNN performs better than LDR-SVM when predicting activation and valence. In contrast, LDR-SVM performs better when predicting emotion. Similar observations can be perceived on corresponding baselines. These experimental results may indicate that, the proposed latent discriminative representation learning method is plausible

Table I. Performance comparison (F1 and UAR) between the proposed Latent Discriminative Representations (LDR) and traditional hand-crafted features on the *dev*elopment and the *test* partitions, by using SVM or RNN for *activation*, *valence*, and *emotion* predictions, respectively, based on audio signals. Results that obtain the best performance are highlighted.

		dev		test	
[%]	methods	F1	UAR	F1	UAR
activation	SVM	57.4	52.1	52.0	50.5
	LDR-SVM	58.9	53.8	56.3	54.6
	RNN	57.6	53.2	57.8	53.1
	LDR-RNN	**59.5**	**55.0**	**58.5**	**56.3**
valence	SVM	54.2	52.9	50.5	49.8
	LDR-SVM	56.7	55.1	52.8	51.6
	RNN	**57.0**	**55.3**	51.8	51.0
	LDR-RNN	56.6	55.2	**53.6**	**52.5**
emotion	SVM	57.8	60.3	51.7	55.0
	LDR-SVM	**59.8**	**61.5**	**53.9**	**56.5**
	RNN	56.2	59.0	50.7	54.5
	LDR-RNN	56.8	58.9	52.6	55.2

to promote performances further when an appropriate classifier is firstly selected for the task at hand.

5. Conclusion

We have presented a latent discriminative representation learning framework to learn discriminative feature representations for speech emotion classification. In this framework, training samples are randomly grouped into positive and negative pairs based on the class labels, and then projected into a latent representation space via a network with the objective of increasing the distance of negative pairs and decreasing that of positive ones. We provided exhaustive experiments to assess the effectiveness and robustness of the proposed framework. In the future, we plan to adjust the method for the task of emotion regression, whereas how to group the training instances with continuous labelling needs to be addressed. Furthermore, we plan to explore the proposed representation learning strategy for other acoustic tasks such as soundscape quality assessment [12] and bird sound classification [13].

Acknowledgement

This work was supported by the European Union's Horizon 2020 Programme through the Innovative Action No. 645094 (SEWA) and a TransAtlantic Platform "Digging into Data" collaboration grant (ACLEW) with the support of the UK's ESRC through the research Grant No. HJ-253479.

References

[1] R. Maskeliunas, V. Raudonis, R. Damasevicius: Recognition of emotional vocalizations of canine. Acta Acust united Ac **104** (2018) 304–314.

[2] M. Rajković, S. Jovičić, D. Grozdić, S. Zdravković, M. Subotić: A note on acoustic features in pitch contours for discrimination of happiness and anger. Acta Acust united Ac **104** (2018) 369–372.

[3] P. Huang, X. He, J. Gao, L. Deng, A. Acero, L. Heck: Learning deep structured semantic models for web search using clickthrough data. Proc. CIKM, San Francisco, CA, 2013, 2333–2338.

[4] E. Hoffer, N. Ailon: Deep metric learning using triplet network. International Workshop on Similarity-Based Pattern Recognition, Copenhagen, Denmark, 2015, 84–92.

[5] G. Le Lan, D. Charlet, A. Larcher, S. Meignier: A triplet ranking-based neural network for speaker diarization and linking. Proc. INTERSPEECH, Stockholm, Sweden, 2017, 3572–3576.

[6] H. Bredin: Tristounet: triplet loss for speaker turn embedding. Proc. ICASSP, New Orleans, LA, 2017, 5430–5434.

[7] J. Han, Z. Zhang, N. Cummins, F. Ringeval, B. Schuller: Strength modelling for real-world automatic continuous affect recognition from audiovisual signals. Image Vision Computing **65** (2017) 76–86.

[8] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, S. Narayanan: IEMOCAP: Interactive emotional dyadic motion capture database. Lang Resour Eval **42** (2008) 335–359.

[9] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, K. Truong: The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE T Affect Comput **7** (2016) 190–202.

[10] Y. Kim, E. M. Provost: Leveraging inter-rater agreement for audio-visual emotion recognition. Proc. ACII, Xi'an, China, 2015, 553–559.

[11] S. Mirsamadi, E. Barsoum, C. Zhang: Automatic speech emotion recognition using recurrent neural networks with local attention. Proc. ICASSP, New Orleans, LA, 2017, 2227–2231.

[12] M. Boes, K. Filipan, B. De Coensel, D. Botteldooren: Machine listening for park soundscape quality assessment. Acta Acust united Ac **104** (2018) 121–130.

[13] K. Qian, Z. Zhang, A. Baird, B. Schuller: Active learning for bird sounds classification. Acta Acust united Ac **103** (2017) 361–364.