Towards Human-Like Holistic Machine Perception of Speaker States and Traits

Yue Zhang, Yifan Liu, and Björn W. Schuller*

Department of Computing, Imperial College London, London, U.K.

Abstract. In this work, we advocate the usage of multi-task deep neural networks with shared hidden layers for various paralinguistic tasks. To this end, the feature transformations are shared across different tasks, while the softmax layers are separately associated with each target label. As a new milestone in holistic speech processing, we constructed a multi-label database in twelve target dimensions, thus enabling large-scale data aggregation for better recognition performance.

1 Introduction

Humans are naturally able to recognise and classify their dialogue partners according to various speaker characteristics from acoustic and linguistic features. Taking humans as a model, the research field of Computational Paralinguistics [1] centers on computer-based recognition of speech phenomena as carried over the voice, ranging from short- through medium-term speaker states such as affect, sleepiness and intoxication, to long-term speaker traits like personality and biological primitives (e. g., age, gender, height, weight). To enable holistic speech processing, the first hurdle to overcome is the general scarcity of multi-label databases, in which instances are annotated in multiple target dimensions. Compounded with the issue of label scarcity, one major shortcoming in current research is that there is very little exploitation of the interrelations between different speaker characteristics, yet in reality, strong interdependencies between bits of paralinguistic information exist.

In our recent works, we were able to verify existent synergies between specific tasks, i. e., deception and sincerity detection [2], non-native prosody score regression and native language classification for language proficiency assessment [3]. In particular, we demonstrated that multi-task learning (MTL) based on data aggregation and classifier chains of auxiliary attributes yields superior performance over single-task learning [2]. To this end, we proposed the Cross-Task Labelling (CTL) method based on Semi-Supervised Learning (SSL) techniques. Moreover, we successfully applied multi-task shared-hidden-layer deep neural networks (MT-SHL-DNN) to emotion recognition, achieving significant better accuracy over single-task DNNs trained with only one emotion representation.

* The research work has received funding from the European Union's Horizon 2020 research and innovation programme under grant No 645378 (ARIA-VALUSPA).

To extend our previous work, we apply the MT-SHL-DNN method to all classification tasks of the Computational Paralinguistic Challenge (ComParE) series for large-scale data and label enrichment. As a result, we achieved a new milestone in this field by creating a holistic multi-label database whose feature vectors comprise the union of instances from all tasks, and whose labels are defined for all instances in all dimensions.

2 Multi-Task Shared-Hidden-Layer DNN

Figure 1a) depicts the architecture of the MT-SHL-DNN, in which the acoustic input features are transformed through the hidden layers shared across different tasks, while each output layer is assigned to a specific target label, with the number of nodes corresponding to the number of classes. The motivation for using this network structure is two-fold: First, multi-task learning acts as a regularisation for the network training, since the hidden layer representation is coerced to be predictive for multiple tasks. Second, it allows an utterance to be interpreted in manifold ways according to various speaker states and traits. The proposed multi-task network structure is far more efficient than using a set of single-task networks, since the input-to-hidden and hidden-to-hidden connections have to be computed only once for each input vector, and the number of parameters in each output layer is small. The fine-tuning of the MT-SHL-DNN is done via error backpropagation and stochastic gradient descent (SGD), where only the shared hidden layers and the task-specific output layer are updated.

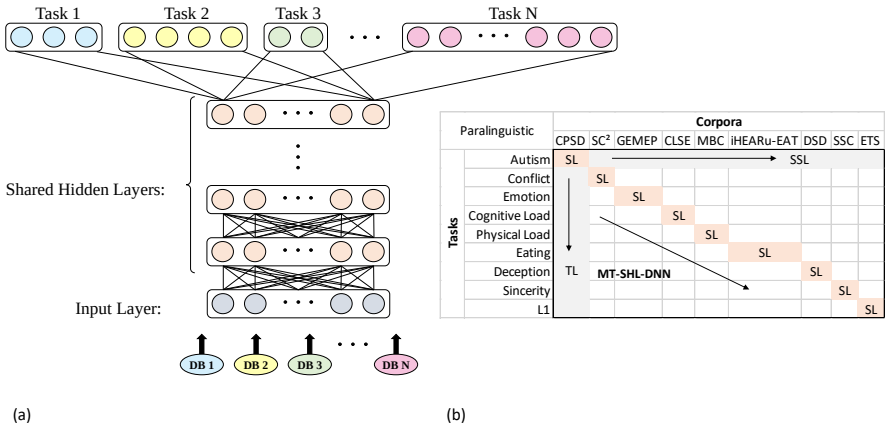

Fig. 1. (a) Structure of the multi-task shared-hidden-layer deep neural network (MT-SHL-DNN), and (b) Holistic matrix for paralinguistic tasks.

3 Application to Paralinguistic Classification Tasks

The INTERSPEECH Computational Paralinguistics ChallengE (ComParE) provides the first of its kind unified test-bed for a broad range of paralinguistic recognition tasks. In this work, we focus on the classification tasks from all ComParE instalments since 2013. The sub-challenges have been carried out on various databases, in detail the Child Pathological Speech Database (CPSD), SSPNet Conflict Corpus (SC^2), the Geneva Multimodal Emotion Portrayals (GEMEP), Cognitive Load with Speech and EGG (CLSE), Munich Biovoice Corpus (MBC), iHEARu-EAT, Deception Speech Database (DSD), Sincerity Speech Corpus (SSC), and ETS Corpus of Non-Native Spoken English. For feature extraction, we use the *ComParE* set containing 6 373 static features resulting from the computation of various functionals over low-level descriptor (LLD) contours. The network topology is formed by four hidden layers of 2 048 neurons each. For the training process, the learning rate was set to 0.01 and Nesterov momentum with coefficient 0.9 was used. To construct our holistic paralinguistic database, we forwarded all training instances of the ComParE corpora through the MT-SHL-DNN using all output layers, thereby obtaining the desired labelling of all data in terms of universal speaker characteristic.

4 Conclusion

In this work, we applied the MT-SHL-DNN method to all ComParE classification tasks, with the aim to label the databases in all twelve target dimensions, thus enabling holistic analysis of speaker characteristics. In Figure 1b), we visualise this paradigm as a *holistic matrix*. There, conventional supervised learning (SL) can be presented by single diagonal elements, while transfer learning (TL) covers the columns. Semi-supervised learning (SSL), as in the rows of the matrix, can be used to predict missing target labels for data aggregation, as e. g., implemented in Cross-Task-Labelling [2]. Corresponding to the diagonal of the holistic matrix, the MT-SHL-DNN method can be understood as using all corpora to learn all tasks at the same time. Finally, by considering all elements of the matrix, we aim to achieve human-like holistic machine understanding of speaker characteristics.

References

1. B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Wiley, 2013.
2. Y. Zhang, F. Weninger, Z. Ren, and B. Schuller, "Sincerity and deception in speech: Two sides of the same coin? A transfer- and multi-task learning perspective," in *Proc. of Interspeech*, (San Francisco, CA), pp. 2041–2045, ISCA, 2016.
3. Y. Zhang, F. Weninger, A. Batliner, F. Hönig, and B. Schuller, "Language proficiency assessment of English L2 speakers based on joint analysis of prosody and native language," in *Proc. of ICMI*, (Tokyo, Japan), ACM, 2016. 5 pages, to appear.