A Deep Learning Framework for Hybrid Linguistic-Paralinguistic Speech Systems

Haytham M. Fayek

School of Engineering, RMIT University
Melbourne, VIC 3001, Australia
haytham.fayek@ieee.org

Abstract

There are two aspects to speech: the linguistic aspect and the paralinguistic aspect. Hybrid linguistic-paralinguistic speech systems that employ joint knowledge of both aspects may fundamentally contribute to better speech processing and understanding, and improve the recognition of both the linguistic and paralinguistic aspects. However, formulating such a hybrid system is a non-trivial task due to unsolved challenges in each of its constituting systems as well as challenges that arise from attempting to integrate such systems. We provide a framework to assess the feasibility of the proposed integration using transfer learning and propose a formulation to hybrid linguistic-paralinguistic speech systems using deep learning.

Index Terms: deep learning, emotion recognition, hybrid systems, neural networks, speech recognition

1. Introduction

Speech conveys linguistic information together with paralinguistic states and characteristics [1]. Knowledge of both aspects of speech—the linguistic aspect and the paralinguistic aspect—is key to better speech processing and understanding and it can be argued that the knowledge of either may better facilitate modeling and classification of the other. For instance, knowledge of the linguistic aspect of speech may boost the detection of the paralinguistic state, and conversely, knowledge of the paralinguistic characteristics may aid in characterizing the linguistic aspect.

Let Automatic Speech Recognition (ASR) be the linguistic task and Speech Emotion Recognition (SER) be the paralinguistic task. For the ASR task, the acoustic model usually utilizes a few frames to recognize phonemes that are later decoded into a transcription, whereas for the SER task, the acoustic model requires a larger number of frames to recognize emotions [2]. Most of the work carried out on ASR considers the presence of emotions in speech a form of distortion and it has been shown generally that the presence of emotions in speech has a negative effect on the accuracy of ASR systems. On the other hand, several studies have reported improvement in SER accuracy when linguistic input is added to the acoustic input [3]. Therefore, it would be advantageous to integrate both systems as sketched in Figure 1, with the aim of improving ASR systems in dealing with emotional speech and at the same time providing linguistic input to SER systems. By extension, hybrid linguistic-paralinguistic speech systems may fundamentally contribute to better speech processing and understanding, and improve the recognition of both the linguistic and paralinguistic components.

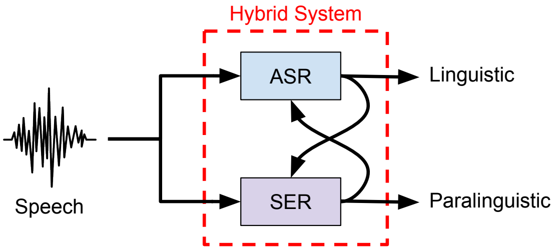

Figure 1: *A Hybrid Linguistic-Paralinguistic System.*

With deep learning being the state-of-the-art approach in ASR [4] and SER, a hybrid system based on deep learning may be viable. However, formulating such a hybrid system is a non-trivial task due to unsolved challenges in each of its constituting systems as well as challenges that arise from attempting to integrate such systems. Thus the aim of this work is to address these challenges and formulate a framework for hybrid linguistic-paralinguistic systems.

2. Challenges

Challenges in ASR are well-documented [5], such as environment noise, speaker attributes, domain adaptation and so on. Fortunately, these challenges have been addressed and mitigated with relative success in recent years. On the other hand, there is a profusion of challenges in SER, where some challenges are inherited from psychology such as the definition of emotions and their representation; and others arise from engineering [6]; such as designing relevant features and dealing with the subjectiveness of the task. In our previous work, we attempted to address some of these challenges: e.g. in [2], we proposed mimicking the ASR pipeline by learning a mapping from Fourier-transform based filter banks to emotion classes using end-to-end deep learning as opposed to feature engineering and demonstrated the feasibility of the proposed scheme; in [7], we proposed modeling the subjectiveness in emotion recognition using soft labels generated from multiple annotators resulting in a richer representation of the underlying emotions which yielded improved performance compared to ground truth labels obtained by majority voting between the same annotators.

Challenges that may arise from the integration of both tasks include choosing an appropriate model architecture, a corresponding learning procedure, and mitigating potential computational complexity. As discussed in Section 3, these challenges may be addressed using a multi task learning framework coupled with an attention mechanism.

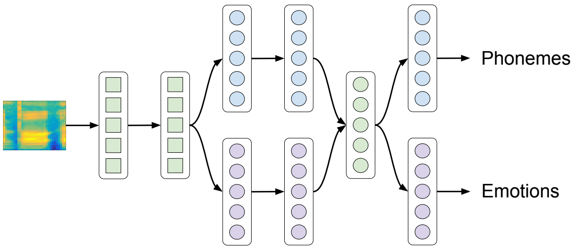

Figure 2: *A Hybrid ASR-SER System. A spectrogram is fed into shared convolutional layers which are followed by a number of specialized layers that share a layer or more to communicate between one another.*

3. Approach

Prior to formulating a hybrid system, the correlation between the linguistic task and the paralinguistic task must first be studied to asses the feasibility of the proposed integration. Transfer learning is a form of knowledge transfer [8], where a model is trained on a base task and data set and then the trained model is repurposed to another target task or data set. Using transfer learning, the correlation between both tasks can be quantified by studying the transferability and relevance of features learned for one task to the other. Assuming that if the transferred features (without fine-tuning) are relevant to the target task, an insignificant or no drop in performance would be observed and conversely, a drop in performance would be observed if the transferred features are irrelevant to the target task; one can then infer a correlation between both tasks by systematically varying the portion of transferred features and measuring the difference in performance, as proposed in our recent work [9].

Once a correlation between both tasks has been established, there can be numerous plausible formulations for a hybrid linguistic-paralinguistic system. Figure 1 depicts a schematic of the general formulation of such a system, where both ASR—the linguistic task—and SER—the paralinguistic task—use minimal speech processing such as Fourier-transform based filter banks and a deep multi-layered model to learn multiple levels of features to map input speech into phonemes or emotions respectively. It is convenient to have a similar architecture in both systems to facilitate their integration.

Transfer learning can also be used to embed knowledge learned from one task into the other by using features learned for one task as an initialization to the other task possibly with a number of fixed layers [9, 10]. Embedding knowledge in a model through transfer learning does not however provide a means for interaction between models, hence a more complex architecture is required. A multi-task learning [11] framework, where several models with shared representations are trained for related tasks, could be employed to achieve both tasks—the linguistic task and the paralinguistic task—in a unified framework to leverage knowledge learned for each task. This could be achieved by choosing an appropriate network architecture that utilizes several shared layers (e.g. convolutional layers) and eventually a number of specialized layers for each task as shown in Figure 2. This concept could be augmented with attention mechanisms [12] to learn a parametric function to weigh the influence of each task to one another.

4. Conclusions

Hybrid linguistic-paralinguistic systems are a potential approach to better speech processing and understanding. Formulating a hybrid linguistic-paralinguistic speech system is a non-trivial task due to unsolved challenges in its constituting systems as well as challenges that may arise from integrating such systems. Recent work has demonstrated the feasibility of such hybrid system by studying the correlation and relevance of features learned between its constituting systems through transfer learning. Future work comprises developing a multi task learning framework with an incorporated attention mechanism.

5. Acknowledgments

This research was funded by the Vice-Chancellor's PhD Scholarship (VCPS) from RMIT University. The author is grateful to Margaret Lech and Lawrence Cavedon for the insightful comments and helpful discussions.

6. References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, Jan 2001.

[2] H. M. Fayek, M. Lech, and L. Cavedon, "Towards real-time speech emotion recognition using deep neural networks," in *Signal Processing and Communication Systems (ICSPCS), 2015 9th International Conference on*, December 2015.

[3] C. M. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 2, pp. 293–303, March 2005.

[4] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[5] X. Huang, J. Baker, and R. Reddy, "A historical perspective of speech recognition," *Commun. ACM*, vol. 57, no. 1, pp. 94–103, Jan. 2014.

[6] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 910, pp. 1062 – 1087, 2011.

[7] H. M. Fayek, M. Lech, and L. Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *2016 International Joint Conference on Neural Networks (IJCNN)*, July 2016.

[8] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," *JMLR: Proc. Unsupervised and Transfer Learning challenge and workshop*, pp. 17–36, 2012.

[9] H. M. Fayek, M. Lech, and L. Cavedon, "On the correlation and transferability of features between automatic speech recognition and speech emotion recognition," in *Interspeech*, September 2016.

[10] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3320–3328.

[11] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 7 1997.

[12] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *CoRR*, vol. abs/1506.07503, 2015.