Deep End-to-End Representation Learning for Food Type Recognition from Speech

Benjamin Sertolli
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
benjamin.sertolli@student.uni-augsburg.de

Nicholas Cummins
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
nicholas.cummins@ieee.org

Abdulkadir Sengur
Firat University, Technology Faculty, Electrical & Electronics Engineering Department, Elazig, Turkey
ksengur@firat.edu.tr

Björn W. Schuller*
GLAM – Group on Language, Audio & Music, Imperial College London, UK
schuller@ieee.org

ABSTRACT

The use of Convolutional Neural Networks (CNN) pre-trained for a particular task, as a feature extractor for an alternate task, is a standard practice in many image classification paradigms. However, to date there have been comparatively few works exploring this technique for speech classification tasks. Herein, we utilise a pre-trained end-to-end Automatic Speech Recognition CNN as a feature extractor for the task of food-type recognition from speech. Furthermore, we also explore the benefits of Compact Bilinear Pooling for combining multiple feature representations extracted from the CNN. Key results presented indicate the suitability of this approach. When combined with a Recurrent Neural Network classifier, our strongest system achieves, for a seven-class food-type classification task an unweighted average recall of 73.3 % on the test set of the iHEARu-EAT database.

CCS CONCEPTS

• **Information systems → Speech / audio search**; • **Applied computing → Health informatics**; • **Computing methodologies →** Speech recognition; Learning latent representations;

KEYWORDS

Eating Condition, Deep Representation Learning, End-to-End Learning, Compact Bilinear Pooling, Recurrent Neural Networks

ACM Reference Format:
Benjamin Sertolli, Nicholas Cummins, Abdulkadir Sengur, and Björn W. Schuller. 2018. Deep End-to-End Representation Learning for Food Type Recognition from Speech. In *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3242969.3243683

*Björn Schuller is also with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany.

ICMI'18, October 16–20, 2018, Boulder, CO, USA

ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3243683

1 INTRODUCTION

The World Health Organisation recently estimated that global obesity levels have tripled since 1975 [38]. Obesity places a strain on health services, as it is associated with an increased risk for a range of serious health conditions such as cardiovascular disease, hypertension and type 2 diabetes [27]. Furthermore, this increase has led to a corresponding increase in the economic burden on society; it was estimated that in 2014 the total global economic impact of obesity was $2.0 trillion (US dollars) [11]. Therefore, there is an urgent and increasing need to reduce obesity levels, both as a public health and economic responsibility [36].

In this regard, there are increasing research efforts into robust and reliable techniques for automatically monitoring food intake e. g. , [3, 5, 13]. The majority of these solutions are based on wearable technologies whose reliability is easily affected by issues relating to adherence and usability [3]. The food-type sub-challenge in the 2018 *Eating Analysis & Tracking* (EAT) Challenge, on the other hand, is focused on encouraging audio and/or visual based approaches for the task of recognising what food type a person is eating whilst speaking [18]. With the increasing ubiquity of smartphones and voice-activated personal assistants [29], remote audio-visual based solutions offer a promising approach for monitoring food intake.

The effect that food-type has on speech has been previously studied using the iHEARu-EAT database [19], as part of the Interspeech 2015 *Computational Paralinguistics Challenge* (ComParE) [34] which also focused on the same 7-class problem as the EAT food type sub-challenge. Within that challenge, Milde and Biemann [30] utilised a *Convolutional Neural Network* (CNN) in a representation learning approach. More specifically, the authors trained a logistic regression classifier on activations from a CNN classifier pre-trained for a 7-class language classification problem. Despite the lack of obvious relationship between the language and food-type problems, this approach still obtained strong accuracies [30].

Outside of the challenge, Hantke et al. [19], explored the effects that eating while speaking had on an *Automatic Speech Recognition* (ASR) system. Results presented by the authors indicated statistically significant increases in *Word Error Rate* (WER) and *Character Error Rate* (CER) when speakers were eating compared to when they were speaking normally without food. Further, the obtained WER and CER errors had statistically significant differences in their distributions across the different food-types tested.

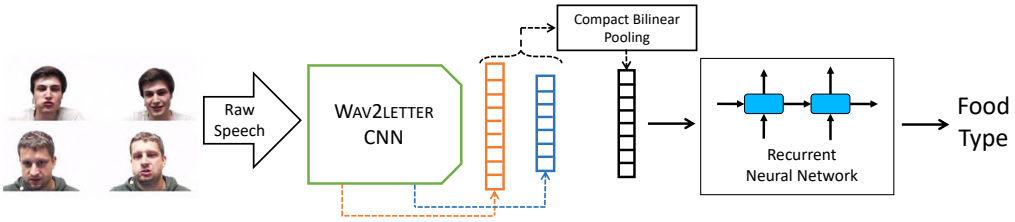

Figure 1: An illustrative overview of the proposed system for speech-based food type recognition. First, raw audio samples from the iHEARu-EAT database are fed into the wav2letter convolutional neural network. Feature vectors are then formed by extracting the activations of different network layers and combining them using compact bilinear pooling. Finally, a recurrent neural network is used to classify the speech into one of seven classes depending on the food type being eaten. Note, the image of the participants in the iHEARu-EAT database (left side) has been reproduced from [19].

Motivated by the results presented in both [19, 30], we herein present our contribution to the food-type (EAT) sub-challenge which utilises the pre-trained wav2letter [8, 28] CNN for feature representation learning. The wav2letter CNN is an end-to-end network for ASR which, to the best of the authors' knowledge, has yet to be used for representation learning. CNN based representation learning has started to be explored as a feature extraction technique for computational paralinguistic due to its ability to provide robust, and often task-specific, features [2, 4, 9, 26, 37].

Further, when using CNNs for representation learning, it is well known that the different layers provide a hierarchy of feature representation, from broad in the initial layers to more task specific in the later layers. Therefore, as well as testing individual layers from the wav2letter CNN, we also combine information using *compact bilinear pooling* (CBP) [14, 15]. CBP has previously been used in computational paralinguistics task e. g. , [1, 40], and in other recognition tasks such as image classification [15, 24], visual question answering [14, 39], and Neural Machine Translation [10, 23].

The rest of this paper is laid out as follows; first the proposed methodology, including the wav2letter CNN, compact bilinear modelling and our proposed classification approach are outlined in Section 2. Then, the key experimental settings are given in Section 3 with the subsequent results being given in Section 4. Finally, we conclude the paper and offer future work directions in Section 5.

2 PROPOSED METHODOLOGY

The proposed approach for the recognition of what food type a person is eating whilst speaking (cf. Figure 1) utilises a pipeline of feature representation learning (cf. Section 2.1), compact bilinear pooling (cf. Section 2.2), and a recurrent neural network classifier (cf. Section 2.3).

2.1 End-to-End Speech Recognition Engine

While there are an abundance of pre-trained (deep) CNN models available for a range of image-based classification tasks, the availability of such CNNs for automatic speech recognition tasks is comparably small [16]. Recently, the Facebook AI Reseach Group released their wav2letter ASR system [8, 28]. This system is an end-to-end approach and provides a model, pre-trained on the Librispeech dataset [31], which takes in a raw speech signal as an input, and outputs a text transcription. Furthermore, the wav2letter approach is trained using graphemes which negates the need for any forced alignment within the classification pipeline.

It is worth noting here that within wav2letter there are a total of 61 'steps' involved between feeding in the raw speech to obtaining the output letter scores. Given that, to the best of the authors' knowledge, this is the first time a wav2letter CNN has been used for representation learning, a key hyperparameter to investigate during system development will therefore be which layer(s) produce useful feature representations for the task at hand.

2.2 Compact Bilinear Pooling

The features elicited from the wav2letter network should vary in level of abstraction depending exactly from which layer they are extracted. We therefore speculate that combining information from different layers will potentially be beneficial to our classification system. In this regard, we explore the benefits of *Compact Bilinear Pooling* (CBP) to combine information from different layers [15].

The *bilinear pooling* operation [35], constructs a global feature vector for a given input image, using the outer product of two vectors $x \in \mathbb{R}^{n_1}$ and $y \in \mathbb{R}^{n_2}$:

$$z = W\,[x \otimes y]\,, \tag{1}$$

where W denotes the learnt linear model, $\otimes$ the outer product operation, and [] the linearisation of the outer product matrix.

The advantage of using the outer product is that it allows for the multiplicative interaction between *all* elements of both vectors [35]. However, this leads to both high memory cost and long computation time. The CBP transform was proposed to reduce the complexity associated with bilinear pooling [15]. This is achieved by creating a more compact feature representation via a low dimensional projection function; i. e. , $\phi(x) \in \mathbb{R}^d$, where $d \ll n$.

In this work, we use the Count Sketch projection function[1]. Given an input vector, $x \in R^n$ which is to be projected into the output vector $x' \in R^d$, the Count Sketch methodology first initialises two random vectors s and h, both $\in R^d$, from a uniform distribution. All elements in s are $\in \{1, -1\}$; the role h on the other hand, is to map between the input and output dimensionality; therefore, all entries in h are $\in \{1, \ldots, d\}$; after their initial assignment, the entries of both vectors are held constant for all applications of the transform. x' is then initialised as a zero vector. Finally, for every element $x[i]$ its destination index $j = h[i]$ is looked up using h, and $s[i] \cdot x[i]$ is added to $x'[j]$.

[1] https://github.com/therne/compact-bilinear-pooling-tf

A further advantage of the Count Sketch projection is that the outer product pooling operation is not explicitly performed. Instead, it is reworked as a convolution operation [33], such that:

$$(x \otimes y, h, s) = (x, h, s) \circledast (y, h, s), \tag{2}$$

where $\circledast$ denotes the convolution operator. As convolution is an element-wise operation in the frequency domain, the CBP transform can be implemented in a computationally efficient manner using the *Fast Fourier Transform* (FFT):

$$z = FFT^{-1}(FFT(x') \odot FFT(y')), \tag{3}$$

where $\odot$ denotes the element-wise product operand, and x' and y' the reduced version of input vectors x and y respectively.

2.3 Classification

Recurrent Neural Network (RNN) classification is utilised to capture any food-type specific temporal modulations in the associated speech files. RNNs, as opposed to other neural network topologies, contain cyclical connections endowing them with the capability of accessing previously processed inputs. To overcome the vanishing gradient problems associated with this paradigm [6], we implement *Long-Short-Term-Memory* (LSTM) neurons [20]. The suitability of this classification paradigm for computational paralinguistic tasks can be seen in the relevant literature, e. g. , [7, 12, 17, 21, 25, 37].

3 EXPERIMENTAL SETTING

All key experimental settings associated with the presented results are outlined in the following subsections.

3.1 Challenge Corpus and Conditions

All experimental results are presented on the iHEARu-EAT database [18, 19, 34]. This corpus contains audio-visual records of 30 subjects speaking either without food in their mouth, or speaking whilst eating with one of six different food types (cf. Table 1). For full details on the make-up of the dataset and the associated recording paradigm, the interested reader is referred to [19].

For the EAT challenge [18], the data was divided into a training set (20 speakers) and test set (10 speakers). As per the challenge guidelines, all results on the training set are reported in terms of *Unweighted Average Recall* (UAR) found using *leave-one-speaker-out cross-validation* (LOSO-CV). The test set results are reported in terms of UAR calculated using the food-type predictions from the 10 speakers contained.

3.2 General System Settings

The Wav2letter CNN is implemented in Torch, specially, we utilise the pre-trained *librispeech-glu-highdropout* model[2]. Wav2letter operates on wav-files sampled at 16 kHz. The output scores are produced using a window of 1955 ms, with steps of 20 ms. To reduce the amount of training data being fed into our RNN classifiers, we implemented an average-pooling methodology based on dividing the audio files into chunks for every started 5-second interval; e. g. , a 10-second file is divided into 2 output chunks, while a 12-second file is divided (evenly) into 3 frames.

[2]https://github.com/facebookresearch/wav2letter

Table 1: An overview of the distribution of the number of utterances per class in the training partition of the iHEARu-EAT dataset, also given is the mean and standard deviation (Std. Dev.) file length within each class. As the class distribution of the iHEARu-EAT test set is not publicly available, only the total number of utterances, including the relevant timing information, is given for this partition.

	Number of Utterances	File Length (Sec) Mean	Std. Dev.
Training Partition			
No Food	140	6.6	2.6
Apple	140	7.3	2.8
Banana	140	7.0	2.7
Biscuit	133	7.5	3.0
Crisp	140	7.2	2.7
Gummi Bear	119	7.0	2.7
Nectarine	133	7.3	2.8
Test Partition	469	7.8	2.7

Initial testing of the extracted feature vector was conducted using a Support Vector Machine (SVM). We implemented a Linear-SVM using Scikit-learn [32], and tested complexity values (C) in the range $\{2^{-7}, ..., 2^{1}\}$ using a step size of 1 in the exponent. Majority voting was used to produce a single prediction per audio file. During system development, we implemented two different RNN classifiers in Tensorflow (v1.8.0 for GPU). The first, herein denoted as RNN-1, was a uni-directional 2-layer LSTM with 256 units per layer. This network was trained using Adam optimisation, a learning rate of 0.001, a batch size of 128, no batch normalisation and no dropout. The second RNN, herein denoted as RNN-2, was also a uni-directional 2-layer LSTM, however, with 192 units per layer. This network was trained using Stochastic Gradient Descent (SGD) optimisation with an exponential decay learning rate of $0.02 \times 0.95^{epoch\_number}$ and a momentum of 0.9. RNN-2 also had a batch size of 128, with batch normalisation and a dropout rate of 0.1. All networks had a final fully-connected output layer with one unit for each predictable class; for classification the unit with the largest output value was considered the predicted class.

3.3 Settings for Test Submissions

In keeping with the official EAT guidelines, we submitted 5 test results for submission. Our test-set submissions are all based on fusing information from different layers using CBP and a RNN classifier. In our first grouping of submissions we utilised CBP to combine features from wav2letter layers 10 and 13 in a 1 000 dimensional representation. For Test Set-up 1, the RNN was a 2-layer LSTM with 256 units trained using Adam optimisation, a learning rate of 10^{-5}, a batch size of 128, batch normalisation and 400 Epochs. Test Set-up 2 was based on RNN-2, however, without dropout and 50 Epochs. Test Set-up 3 was also based on RNN-2 – this time with no dropout, a momentum setting of 0.5, a learning rate of $0.1 \times (0.95^{epoch})$, and 80 Epochs. Our final two entries utilised CBP to combine entries from multiple layers, specifically wav2letter layers 10 and 13 were combined in a 1 000-dimensional representation as were layers 13 and 16. These two representations

Table 2: A comparison of UARs for a 7-class food-type recognition task. The scores are found using LOSO-CV on the training partition of the iHEARu-EAT dataset. Displayed are the results from the five best features extracted from different layers of the wav2letter CNN with either our SVM, including the corresponding complexity value (C), RNN-1 or RNN-2 classification systems. Also displayed are the results obtained by using Compact Bilinear Pooling to combine the extracted features, which are then classified using either our RNN-1 and RNN-2 classifiers.

Layer (Dimensionality)	SVM (C)	RNN-1	RNN-2
Layer 10 (484)	**61.4 %** (2^{-3})	**59.1 %**	73.4 %
Layer 13 (532)	60.5 % (2^{0})	56.9 %	**73.8 %**
Layer 16 (584)	57.7 % (2^{-2})	53.9 %	67.5 %
Layer 19 (642)	53.9 % (2^{1})	51.1 %	64.3 %
Layer 25 (776)	52.8 % (2^{-3})	46.5 %	61.1 %
Pooled Layers (Dimensionality)		**RNN-1**	**RNN-2**
10 + 13 (500)		46.8 %	74.3 %
10 + 13 (1000)		41.8 %	75.4 %
(10 + 13 (1000)) + (13 + 16 (1000)) (1500)		38.7 %	**76.4 %**

were then pooled to form a 1 500-dimensional feature vector. Test Set-up 4 was, again, based on RNN-2 with 100 Epochs of training. Finally, Test Set-up 5 utilised a RNN set-up similar to RNN-2, however, with 0.2 dropout and again 100 Epochs of training.

4 RESULTS AND DISCUSSION

To initially assess the feasibility of using the wav2letter CNN for speech-based representation learning, we conducted a series of tests using a SVM classifier on the iHEARu-EAT training partition (cf. Table 2). These tests indicated the promise of the obtained features with the strongest SVM UAR, 61.4 % being slightly below the EAT audio baseline of 64.4 % set using a *Bag-of-Audio-Words* (BoAW) paradigm [18]. Disappointingly, our initial RNN tests, performed with RNN-1 (Adam Optimiser), did not obtain as strong a performance as when using SVMs (cf. Table 2). However, our strongest results were found when using the RNN-2 (SGD optimiser) set-up. This paradigm achieved a maximum UAR of 73.8 %, well beyond both our SVM system and the challenge baseline.

Interestingly, the results obtained across all our classifiers indicate features extracted from the lower layers of wav2letter are better suited to the task of food type recognition (cf. Table 2). Potentially, the features extracted from the higher layers are too specific to the English speech recognition task the network was originally trained for. Using CBP to fuse features extracted from the different wav2letter layers – in combination with our RNN-2 classifier – improved system performance beyond those achieved for the single layer representation (cf. Table 2). Our strongest training set UAR, 76.4 %, was achieved using a multiple pooling operation to combine information from layers 10, 13 and 16 of the wav2letter CNN.

The challenge organisers provided two official audio baselines [18]: the first, realised with a BoAW and SVM set-up, with a UAR of 67.2 %, and the second, realised with a CNN-RNN based end-to-end paradigm, with a UAR of 32.8 %. All our test-set submissions returned scores over the challenge baselines (cf. Table 3), it is worth noting here that a Kolmogorov-Smirnov test revealed no significant differences between these results. Test Set-Up 2 gained our strongest performance – a UAR of 73.3 %. This result represents a relative improvement of 9.1 % on the BoAW baseline and a 123.5 % improvement over the end-to-end system.

Table 3: A comparison of UARs for a 7-class food-type recognition task. The scores are reported for the train(ing) partition using LOSO-CV, as well as the test partition of the iHEARu-EAT dataset. All systems utilised Compact Bilinear Pooling to combine features extracted from the wav2letter and a RNN based classifier.

Model	Train UAR	Test UAR
Test Set-Up 1	74.1 %	69.4 %
Test Set-Up 2	75.0 %	**73.3 %**
Test Set-Up 3	75.2 %	70.1 %
Test Set-Up 4	76.6 %	71.9 %
Test Set-Up 5	75.8 %	72.6 %

When comparing our test results with notable systems from the 2015 ComParE challenge, our approach was able to achieve comparable performances with the CNN approach presented in [30], which obtained a test set UAR of 75.9 %. However, our scores are below the winning UAR of 83.1 % [22]. This system successfully utilised speaker clustering and repeated normalisation to reduce confounding effects relating to speaker identity, while such effects were not accounted for in our approach. Critically, the strong test set performances gained by our approach show the promise of using wav2letter, especially in combination with CBP, for speech based representation learning.

5 CONCLUSIONS

Convolutional Neural Network (CNN) based representation learning is a common feature extraction technique in many machine learning recognition tasks. Results presented in this paper indicate that the recently released wav2letter CNN can produce useful speech features. Furthermore, Compact Bilinear Pooling (CBP) can be leveraged to construct feature vectors containing information from multiple Wav2letter layers. Results gained on the iHEARu-EAT corpus show that the combination of wav2letter, CBP and a recurrent neural network can gain a 73.3 % UAR on a seven class food type classification problem. Future work will aim at verifying these results on similar computational paralinguistic tasks – in particular, depression, fatigue, and intoxication detection, as effects relating to these conditions have also been manifested at the phonetic structure of the affected utterances.

ACKNOWLEDGMENTS

This work was supported by the European Unions's Seventh Framework and Horizon 2020 Programmes under grant agreement No. 338164 (ERC StG iHEARu).

REFERENCES

[1] Z. Aldeneh, S. Khorram, D. Dimitriadis, and E. M. Provost. 2017. Pooling Acoustic and Lexical Features for the Prediction of Valence. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI 2017)*. ACM, Glasgow, UK, 68–72.

[2] Z. Aldeneh and E. M. Provost. 2017. Using regional saliency for speech emotion recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, New Orleans, Louisiana, 2741–2745.

[3] R. Alharbi, N. Vafaie, K. Liu, K. Moran, G. Ledford, A. Pfammatter, B. Spring, and N. Alshurafa. 2017. Investigating barriers and facilitators to wearable adherence in fine-grained eating detection. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, Kona, Hawaii, 407–412.

[4] S. Amiriparian, M. Gerczuk, S. Ottl, N.s Cummins, M. Freitag, S. Pugachevskiy, and B. Schuller. 2017. Snore Sound Classification Using Image-based Deep Spectrum Features. In *Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*. ISCA, Stockholm, Sweden, 3512–3516.

[5] A. Bedri, R. Li, M. Haynes, R. P. Kosaraju, I. Grover, M. Y. Prioleau, T.and Beh, M. Goel, T. Starner, and G. Abowd. 2017. EarBit: Using Wearable Sensors to Detect Eating Episodes in Unconstrained Environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (Sep. 2017), 37:1–37:20.

[6] Y. Bengio, P. Simard, and P. Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks* 5, 2 (Mar. 1994), 157–166.

[7] R. Brueckner and B. Schulter. 2014. Social signal classification using deep blstm recurrent neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Florence, Italy, 4823–4827.

[8] R. Collobert, C. Puhrsch, and G. Synnaeve. 2016. Wav2Letter: an End-to-End ConvNet-based Speech Recognition System. *CoRR* abs/1609.03193 (2016).

[9] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. Schuller. 2017. An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech. In *Proceedings of the 25th ACM International Conference on Multimedia (MM '17)*. ACM, Mountain View, California, 478–484.

[10] J.-B. Delbrouck and S. Dupont. 2017. Multimodal Compact Bilinear Pooling for Multimodal Neural Machine Translation. *CoRR* abs/1703.08084 (2017).

[11] R. Dobbs, C. Sawers, F. Thompson, J. Manyika, J. R. Woetzel, P. Child, S. McKenna, and A. Spatharou. 2014. Overcoming obesity: an initial economic analysis. https://goo.gl/6R7kz2. Accessed: 31-05-2018.

[12] S. Fernando, V. Sethu, E. Ambikairajah, and J. Epps. 2017. Bidirectional Modelling for Short Duration Language Identification. In *Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*. ISCA, Stockholm, Sweden, 2809–2813.

[13] J. M. Fontana, M. Farooq, and E. Sazonov. 2014. Automatic Ingestion Monitor: A Novel Wearable Device for Monitoring of Ingestive Behavior. *IEEE Transactions on Biomedical Engineering* 61, 6 (June 2014), 1772–1779.

[14] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. *CoRR* abs/1606.01847 (2016).

[15] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. 2016. Compact Bilinear Pooling. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, Nevada, 317–326.

[16] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen. 2018. Recent advances in convolutional neural networks. *Pattern Recognition* 77 (2018), 354–377.

[17] J. Han, Z. Zhang, N. Cummins, F. Ringeval, and B. Schuller. 2016. Strength Modelling for Real-World Automatic Continuous Affect Recognition from Audiovisual Signals. *Image and Vision Computing, Special Issue on Multimodal Sentiment Analysis and Mining in the Wild* 65 (Sep. 2016), 76–86.

[18] S. Hantke, M. Schmitt, P. Tzirakis, and B. Schuller. 2018. EAT – The ICMI 2018 Eating Analysis and Tracking Challenge. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18)*. ACM, Boulder, Colorado. 5 pages.

[19] S. Hantke, F. Weninger, R. Kurle, F. Ringeval, A. Batliner, A. El-Desoky Mousa, and B. Schuller. 2016. I Hear You Eat and Speak: Automatic Recognition of Eating Condition and Food Types, Use-Cases, and Impact on ASR Performance. *PLoS ONE* 11, 5 (May 2016), 1–24.

[20] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[21] C. W. Huang and S. S. Narayanan. 2017. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, Hong Kong, P. R. China, 583–588.

[22] H. Kaya, A. A. Karpov, and A. A. Salah. 2015. Fisher vectors with cascaded normalization for paralinguistic analysis. In *Proceedings INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association*. ISCA, Dresden, Germany, 909–913.

[23] J.-H. Kim, K. W. On, W. Lim, J. Kim, J. Ha, and B.-T. Zhang. 2016. Hadamard Product for Low-rank Bilinear Pooling. *CoRR* abs/1610.04325 (2016).

[24] S. Kong and C. Fowlkes. 2017. Low-Rank Bilinear Pooling for Fine-Grained Classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Honolulu, Hawaii, 365–374.

[25] D. Le, Z. Aldeneh, and E. Mower Provost. 2017. Discretized Continuous Speech Emotion Recognition with Multi-Task Deep Recurrent Neural Network. In *Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*. ISCA, Stockholm, Sweden, 1108–1112.

[26] W. Lim, D. Jang, and T. Lee. 2016. Speech emotion recognition using convolutional and Recurrent Neural Networks. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, Jeju, South Korea, 1–4.

[27] E. A. Lin, G. M. Barlow, and R. Mathur. 2015. *The Health Burden of Obesity*. Springer New York, New York, NY, 19–42.

[28] V. Liptchinsky, G. Synnaeve, and R. Collobert. 2017. Letter-Based Speech Recognition with Gated ConvNets. *CoRR* abs/1712.09444 (2017).

[29] H. Liu, H. Ning, Q. Mu, Y. Zheng, J. Zeng, L. T. Yang, R. Huang, and J. Ma. 2017. A review of the smart world. *Future Generation Computer Systems* (2017). 14 pages, in press.

[30] B. Milde and C. Biemann. 2015. Using representation learning and out-of-domain data for a paralinguistic speech task. In *Proceedings INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association*. ISCA, Dresden, Germany, 904–908.

[31] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brisbane, Australia, 5206–5210.

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

[33] R. Pham, N.and Pagh. 2013. Fast and Scalable Polynomial Kernels via Explicit Feature Maps. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13)*. ACM, Chicago, Illinois, 239–247.

[34] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger. 2015. The INTERSPEECH 2015 Computational Paralinguistics Challenge: Degree of Nativeness, Parkinson's & Eating Condition. In *Proceedings INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association*. ISCA, Dresden, Germany, 478–482.

[35] J. B. Tenenbaum and W. T. Freeman. 2000. Separating Style and Content with Bilinear Models. *Neural Computation* 12, 6 (June 2000), 1247–1283.

[36] M. Tremmel, U.-G. Gerdtham, P. M. Nilsson, and S. Saha. 2017. Economic Burden of Obesity: A Systematic Literature Review. *International Journal of Environmental Research and Public Health* 14, 4 (2017). Article Number 435.

[37] G. Trigeorgis, F. Ringeval, R. Brückner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou. 2016. Adieu Features? End-to-End Speech Emotion Recognition using a Deep Convolutional Recurrent Network. In *Proceedings 41st IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2016*. IEEE, Shanghai, P. R. China, 5200–5204.

[38] World Health Organization (WHO). 2018. Obesity and Overweight. http://www.who.int/mediacentre/factsheets/fs311/en/. Accessed: 26-03-2018.

[39] Z. Yu, J. Yu, J. Fan, and D. Tao. 2017. Multi-Modal Factorized Bilinear Pooling With Co-Attention Learning for Visual Question Answering. In *The IEEE International Conference on Computer Vision (ICCV)*. IEEE, Venice, Italy, 1821–1830.

[40] S. Zhang, S. Zhang, T. Huang, and W. Gao. 2018. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. *IEEE Transactions on Multimedia* 20, 6 (June 2018), 1576–1590.