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

Using the Bag-of-Audio-Word Feature Representation of ASR DNN Posteriors for Paralinguistic Classification

Gábor Gosztolya

MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary
ggabor @ inf.u-szeged.hu

Abstract

The Bag-of-Audio-Word (or BoAW) representation is an utterance-level feature representation approach that was successfully applied in the past in various computational paralinguistic tasks. Here, we extend the BoAW feature extraction process with the use of Deep Neural Networks: first we train a DNN acoustic model on an acoustic dataset consisting of 22 hours of speech for phoneme identification, then we evaluate this DNN on a standard paralinguistic dataset. To construct utterance-level features from the frame-level posterior vectors, we calculate their BoAW representation. We found that this approach can be utilized even on its own, although the results obtained lag behind those of the standard paralinguistic approach, and the optimal size of the extracted feature vectors tends to be large. Our approach, however, can be easily and efficiently combined with the standard paralinguistic one, resulting in the highest Unweighted Average Recall (UAR) score achieved so far for a general paralinguistic dataset.

Index Terms: Bag-of-Audio-Words, Deep Neural Networks, computational paralinguistics, classifier combination

1. Introduction

An emerging area of speech technology is computational paralinguistics, which seeks to extract, locate and identify various types of non-verbal phenomena appearing in human speech. Notable examples of such tasks are emotion detection [1, 2, 3] and laughter localization [4, 5, 6], but numerous other tasks have been investigated such as conflict intensity estimation [7, 8, 9], measuring the amount of cognitive and physical load [10] and various medical applications like detecting Parkinson's disease or Alzheimer's disease and depression [11, 12, 13].

One key aspect in which computational paralinguistics differs from Automatic Speech Recognition (ASR) is that ASR has to handle a variable-size input (i.e. an utterance) and it has to produce a variable-length output (i.e. the transcription). Computational paralinguistics, by contrast, treats one utterance as one example, therefore the output for one utterance has a fixed length (e.g. a class label or some score estimated by regression). However, we still have to process utterances of various lengths. To be able to utilize standard machine learning methods such as classification and regression algorithms, fixed-length feature vectors have to be extracted from these utterances of varying length. These fixed-length feature vectors can then be used for training standard classification algorithms such as Support-Vector Machines (SVMs) [14] and Deep Neural Networks (DNNs).

In the standard solution for utterance-level feature extraction in computational paralinguistics, developed over the years mainly during the Interspeech Computational Paralinguistic ChallengEs (ComParE, see e.g. [15, 16, 17]), first low-level descriptors such as energy, spectral, cepstral (MFCC) and voicing related attributes are computed frame-wise; then these are transformed into utterance-level features by using specific functionals like the mean and standard deviation. This approach also appears to be task-independent, as it has been effectively used in dozens of different tasks [18].

Another technique for utterance-level feature extraction, introduced in the last few years, is that of Bag-of-Audio-Words (BoAW, [19]). In the BoAW approach, inspired by the area of Natural Language Processing and of image processing, we take the frame-level features (e.g. MFCCs or mel filter bank energies) of the utterances of the training set and cluster them. Then, for the next step, each frame-level feature vector is replaced with its cluster; utterance-level feature vectors are calculated as the (normalized) histogram of the clusters of the frame vectors of the given utterance [20]. This process can then be applied for the test set as well: a feature vector unseen during the clustering step (such as those of the test set) can be assigned to one of the already defined clusters based on its Euclidean distance from the cluster centers. In the third step, utterance-level classifier training and evaluation can be realized by utilizing these normalized histograms as utterance-level feature vectors, e.g. by using an SVM.

BoAW representations have been used in various audio processing tasks such as multimedia event classification [20], emotion recognition [21, 22], acoustic event detection [23], snore sound classification [24] and determining whether the actual speaker has a cold [17]. This indicates that the BoAW representation is another, powerful way of extracting features in computational paralinguistics. However, although Pancoast and Akbacak originally proposed the BoAW representation in order to incorporate large, unannotated audio archives into the paralinguistic classification process [20], all the above-listed studies performed BoAW extraction only on the given paralinguistic dataset. Keeping in mind that paralinguistic datasets tend to consist of only a few hours of recordings overall, while in ASR having over a hundred hours is quite common nowadays, making use of such large datasets in the BoAW process might prove to be beneficial. (We will denote the latter dataset the *external* one.)

Incorporating large external corpora into the BoAW process, however, is far from straightforward. The approach which seems to be the easiest would be to extract the frame-level feature vectors from this external dataset and perform the BoAW clustering step on these examples. Unfortunately, a small ASR dataset of a few dozen hours consists of tens of millions of frames, and such a large number of examples is extremely difficult to cluster efficiently. Were we to downsample the frames of the external dataset, we would lose the advantage of utilizing such a large audio source.

In this paper we present an approach that allows us to effectively incorporate such external acoustic corpora for paralinguistic tasks. The key idea is to train a standard DNN acoustic

http://dx.doi.org/10.21437/Interspeech.2019-1163

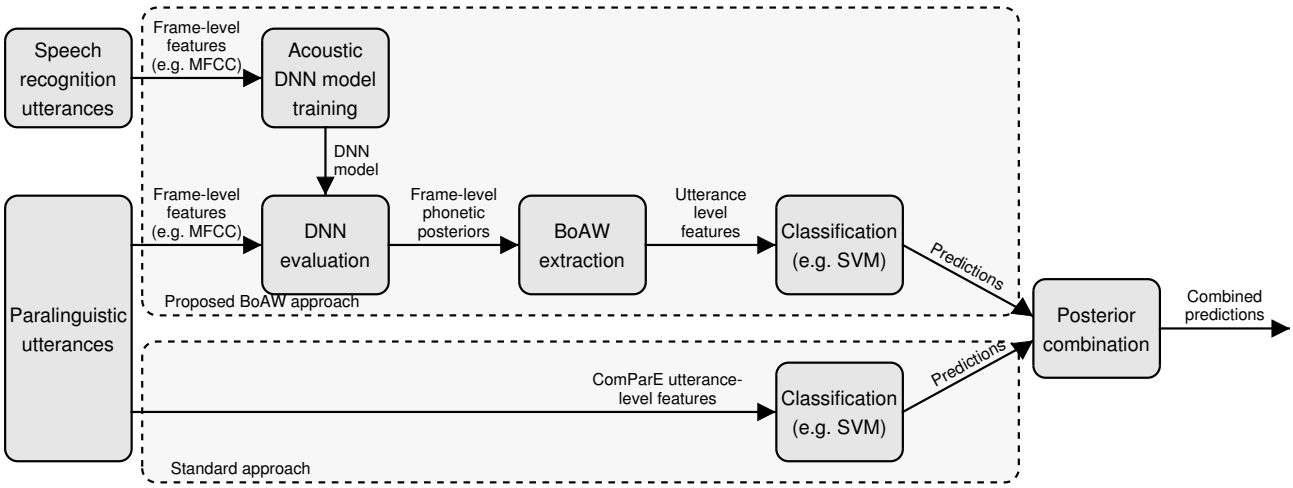

Figure 1: *The general workflow of the proposed BoAW-based classification process.*

model on such an external corpus. In the next step, we evaluate this DNN on the frame-level feature vectors of the given paralinguistic dataset, and perform the BoAW clustering step on the frame-level posterior estimates obtained (i.e. the DNN outputs). Combining the proposed approach with the standard paralinguistic one, we got a 10% improvement in terms of relative error reduction on a public paralinguistic dataset. Further utilizing the BoAW representation of the MFCC feature vectors yielded an even better error reduction score of 32% compared to the baseline, also significantly outperforming the combination of the standard paralinguistic approach with the BoAW-MFCC one. Overall, we report the highest UAR score on a public paralinguistic dataset that was achieved in a scientifically sound manner.

2. BoAW Features of ASR DNN Posteriors

Next, we will describe our proposed workflow. After the description of the general BoAW process, we will explain how we extracted the BoAW representation from the DNN posteriors, and how we fused the predictions of the utterance-level classifier models trained on the various feature sets tested.

2.1. BoAW Representation

The BoAW representation seeks to represent a varying-length utterance by a fixed-length feature vector. Its input is a set of frame-level feature vectors (e.g. MFCCs). In the first step, we cluster these vectors, the number of clusters being a parameter of the method; the resulting cluster centroids will form the *codebook*. Next, each original feature vector is replaced by a single index representing the nearest codeword (*vector quantization*). We calculate the feature vector for the given file by generating a histogram of these indices; and to eliminate the influence of the utterance length, it is common to use some kind of normalization such as the L1 normalization (i.e. divide each cluster count by the number of frames in the given utterance). Furthermore, as an input frame may have a small distance to several audio words (leading to possible ambiguous assignments), we can also take multiple assignments into account.

2.2. BoAW Features of ASR DNN Posteriors

In this study our aim is to incorporate a large audio dataset (i.e. the external one) in the BoAW feature extraction step. In order to do this, first we train a DNN acoustic model on this external corpus. Then, in the next step, we evaluate this DNN on our paralinguistic dataset; BoAW codebook generation and feature extraction is done using these frame-level phonetic posterior estimates. Utterance-level classifier training and evaluation is done by utilizing these BoAW feature representations. (For the general workflow of the proposed approach, see Fig. 1.)

Note that, for the proposed approach, we need corpora that have annotated and time-aligned phonetic labels. In our opinion, however, this is not a limitation from a practical point of view for a number of reasons. Firstly, such datasets are quite easy to obtain. Secondly, they can be re-used for different paralinguistic tasks without the need of re-annotation. Another advantage of relying on such external datasets is that databases used for speech recognition systems tend to be much larger than paralinguistic ones (i.e. hundreds of hours). This increase in dataset size, however, does not hinder the BoAW process, as this only affects the DNN training step.

Although DNN-based ASR systems achieve the best accuracy with the context-dependent (CD) phoneme modelling approach [25], we suggest using context-independent (CI) DNNs instead (where each phoneme is represented by three states, corresponding to the beginning, middle and ending part). We recommend this simplification in order to keep the posterior vector length within a manageable size: ASR systems tend to use thousands or even tens of thousands of context-dependent tied states, which in our case would lead to the same number of DNN outputs, i.e. posterior vector sizes. In contrast, in a CI phoneme state representation we typically have states three times the number of phones (i.e. phoneme beginning, center and ending), usually lying between 100 and 200. Since the posterior scores fall in the unit interval (i.e. $[0, 1]$), we also suggest taking their logarithm first before calculating the codebook vectors.

2.3. Feature Set Combination

Although using the BoAW representation may prove to be beneficial for classification, we should not discard all other kinds of features, especially since the standard paralinguistic feature set

proposed by Schuller et al. proved to be quite effective over the years on several different tasks (see e.g. [15, 17]). Optimality is probably achieved via some combination of the proposed approach with this standard paralinguistic one. In our experience, an effective way of such combining them is to train separate machine learning models for the different types of features, and combine their posterior estimates (called *late fusion*). Now we will take the weighted mean of the resulting posterior scores, which we found to be a simple-yet-robust technique in our earlier studies (see e.g. [26]).

3. Experimental Setup

3.1. The iHEARu-EAT Corpus

We performed our experiments on the *iHEARu-EAT* database [27], which contains the utterances of 30 people recorded while speaking during eating. Six types of food were used along with the "no food" class, resulting in seven classes overall. The total recording time of this dataset is only 2 hours and 51 minutes. For each speaker and food type, seven utterances were recorded; some subjects refused to eat certain types of foods, resulting in a total of 1414 utterances in German. Although this dataset can be used primarily to test machine learning and signal processing techniques, Hantke et al. also anticipated several possible future applications [27]. This dataset was later used in the Interspeech ComParE 2015 Eating Condition Sub-Challenge [15].

3.2. Utterance-level Classification

Our experiments essentially followed the set-up of the ComParE 2015 Challenge [15]. As the standard paralinguistic solution, we used the 6373 ComParE features (see e.g. [15]), extracted by using the openSMILE tool [28]. The feature set includes energy, spectral, cepstral (MFCC) and voicing related low-level descriptors (LLDs), from which specific functionals (such as the mean, standard deviation, percentiles, peak statistics etc.) are computed to provide utterance-level feature values.

We used Support-Vector Machines for utterance-level classification, using the LibSVM [29] library. We applied the nu-SVM method with linear kernel; the value of C was tested in the range $10^{\{-5,\ldots,2\}}$, just like in our previous paralinguistic studies (e.g. [30, 26]). Optimal meta-parameters (e.g. C for SVM or codebook size for BoAW) were determined in speaker-wise cross-validation (CV). To measure performance, we employed the Unweighted Average Recall (UAR) metric. We performed speaker-wise feature standardization, as in our preliminary tests we found that this is beneficial for all feature sets tested; we used the annotated speaker IDs in cross-validation, while the speakers of the test set were identified by using the single Gaussian-based bottom-up Hierarchical Agglomerative Clustering algorithm [31, 32, 33].

3.3. DNN Acoustic Model Training

As the DNN component of the proposed workflow, we applied a deep network, consisting of rectified linear units as hidden neurons [34, 35]. The main advantage of deep rectifier nets is that they can be efficiently trained with the standard backpropagation algorithm, without any tedious pre-training [35, 36]. We used our custom implementation, which achieved the lowest error rate on the TIMIT database ever published with a phonetic error rate of 16.5% on the core test set [37]. We utilized DNNs with three hidden layers, each containing 1000 rectified neurons, and applied the softmax function in the output layer. We used 12 MFCC values along with energy, and their first and second order derivatives [38]; we extended the feature vectors with the features of 7 neighbouring frames from both sides, resulting in a 15 frame-wide sliding window.

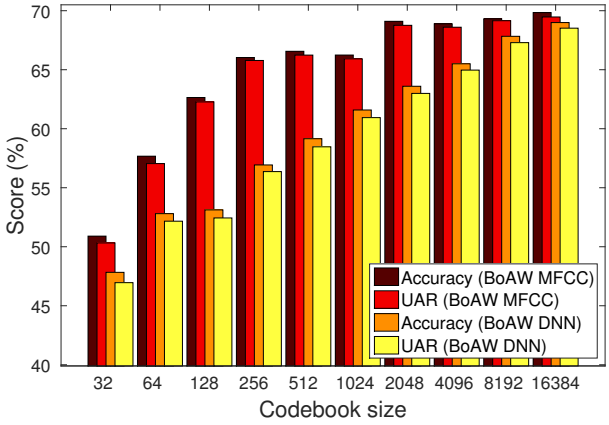

Figure 2: *Classification accuracy and UAR scores as a function of the BoAW codebook size in the cross-validation setup.*

The acoustic models were trained on the "Szeged" Hungarian broadcast new corpus [39]. This 28 hour-long speech corpus was collected from eight TV channels. We then used the 22-hour long training subset for training our CI DNN acoustic model. Since we set up a phonetic set of 51 (Hungarian) phonemes, we had 153 phonetic states, leading to the same number of DNN output neurons. The order of utterances was randomized at the beginning of training. Notice that the paralinguistic dataset had utterances in German, while our external audio database contained Hungarian speech. In our hypothesis, however, this mismatch should not affect the efficiency of our workflow.

3.4. BoAW Parameters

We used the OpenXBOW package [40], which is an open-source toolkit written in Java. We tested codebook sizes of 32, 64, 128, 256, 512, 1024, 2048, 4096 and 8192. We did not increase the codebook size further for two reasons. Firstly, 8192 is already the highest value we found in previous paralinguistic studies (i.e. in [17, 22, 23]). Secondly, having 8192 feature values for each example already leads to a classification process with quite high time and memory requirements, even when the training set size is fairly low.

We used random sampling instead of k-means or k-means++ clustering for codebook generation because it was reported (see [22]) that it provides a similar performance and it is significantly faster. We allowed 5 parallel cluster assignments, i.e. for each frame we chose the 5 closest cluster centers. As we mentioned in Section 2.2, we took the logarithm of the posterior estimates before calculating the codebooks. For comparison, we also tested the standard BoAW approach with the 39-sized MFCC + Δ + $\Delta\Delta$ frame-level feature vectors as input.

4. Results

Figure 2 shows the classification accuracy and UAR scores obtained as a function of the codebook size for the two BoAW approaches tested in the cross-validation setup. It can be seen that

Table 1: *Optimal UAR values for the different feature sets.*

Feature set	CV	Test
ComParE (baseline)	74.3%	74.8%
BoAW-MFCC	69.5%	77.9%
BoAW-DNN	68.5%	66.7%
ComParE + BoAW-MFCC	76.6%	79.5%
ComParE + BoAW-DNN	77.4%	77.3%
BoAW-MFCC + BoAW-DNN	75.2%	79.3%
All three feature sets (proposed)	79.6%	82.8%

increasing the codebook size leads to a higher-quality classification: the highest value tested (16384) produced the optimal accuracy and UAR values for both approaches. Unfortunately, this results in quite large feature vectors, which are difficult to handle. In our future studies, however, it may be worth increasing the codebook size even further.

These models led to competitive scores on the test set as well (see Table 1). In fact, using the MFCC-based BoAW feature representations outperformed the baseline ComParE feature set on the test set, although in CV it lags far behind it, which reflects its non-uniform behaviour. Combining one BoAW approach with the ComParE feature set via late fusion improved the UAR scores further, but we got the best results by combining all three approaches. In this case the weights of the three approaches (determined in the cross-validation setting) were 0.30, 0.35 and 0.35, ComParE, BoAW-MFCC and BoAW-DNN approaches, respectively, which tells us that all three types of feature sets are equally important for achieving state-of-the-art classification accuracy.

Comparing the resulting UAR scores with the other accuracy values published in the literature (see Table 2), the first thing to notice is that our baseline (i.e. using only the ComParE feature set) is significantly higher than the baseline of the ComParE Challenge (65.9%) was [15]. This is mainly due to two reasons. Firstly, we used the libSVM library instead of Weka [41]; and secondly, our baseline already employs speaker-wise feature standardization, which was demonstrated to improve classification accuracy for this particular task (see e.g. [15, 33]).

Furthermore, we can see that the UAR scores attained by applying the proposed method are, in fact, at the same level (78.9%, cross-validation) or even higher (82.8%, test) as Fisher Vector analysis proved to be (see [33]). Also note that the best result reported on this dataset was achieved by guessing combinations of existing techniques, which is a perfectly valid technique in a machine learning challenge, but perhaps not the best approach in a standard scientific study. In contrast, we optimized all meta-parameters (i.e. SVM complexity, BoAW codebook size, late fusion weights) in cross-validation.

The proposed approach, of course, is open to further investigations. For example, it might be worth seeing how the language or the size of the external dataset affects BoAW quality (and therefore, paralinguistic classification as well). The effectiveness of using the BoAW representation of the ASR DNN posteriors might also vary depending on the type of the actual paralinguistic dataset (e.g. variations of emotion detection, identifying speaker stress and conflicts). Moreover, from the technical point of view, the DNN output vectors could be treated as samples from a probability distribution. This observation

Table 2: *Optimal UAR values for the different feature sets.*

Approach	CV	Test
ComParE + BoAW (proposed)	78.9%	82.8%
ComParE 2015 baseline [15]	61.3%	65.9%
Fisher Vectors analysis (Kaya et al., [33])	78.9%	81.6%
Best result reported (Kaya et al., [33])	—	83.1%
Chance	14.3%	14.3%

could also be exploited during the BoAW codebook construction process, for example by using a different distance function. These points, however, fall outside of the scope of the current study, and are the subject of future work.

5. Conclusions

In this study, we sought to incorporate large, unannotated audio datasets for high-quality paralinguistic classification. To this end, we trained context-independent DNN acoustic models on a general speech recognition dataset, and used the frame-level posterior estimates (i.e. DNN outputs) as feature vectors. To construct utterance-level feature values from the frame-level feature vector sets, we employed the Bag-of-Audio-Words approach. According to our experimental results, this technique is viable even on its own, as the UAR values attained were acceptable on a public paralinguistic dataset.

Another possible application of the proposed BoAW-DNN technique is to combine it with other types of features. By combining the predictions got using our method with the classic ComParE functionals proposed by Schuller et al., and also incorporating the predictions obtained by relying on the BoAW representations of MFCC feature vectors, we attained a relative error reduction of about 32%. The UAR score of 82.8% reported is the highest one on this particular paralinguistic corpus achieved so far in a scientifically sound way.

6. Acknowledgements

This study was supported by the National Research, Development and Innovation Office of Hungary via contract NKFIH FK-124413, and by the Ministry of Human Capacities, Hungary (grants 20391-3/2018/FEKUSTRAT and TUDFO/47138-1/2019-ITM). Gábor Gosztolya was also funded by the János Bolyai Scholarship of the Hungarian Academy of Sciences, and by the Hungarian Ministry of Innovation and Technology New National Excellence Program ÚNKP-19-4. The Titan X graphics card used for this study was donated by the NVIDIA Corporation.

7. References

[1] S. L. Tóth, D. Sztahó, and K. Vicsi, "Speech emotion perception by human and machine," in *Proceedings of COST Action*, Patras, Greece, 2012, pp. 213–224.

[2] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The Interspeech 2013 Computational Paralinguistics Challenge: Social signals, Conflict, Emotion, Autism," in *Proceedings of Interspeech*, Lyon, France, Aug 2013, pp. 148–152.

[3] H. Kaya and A. A. Karpov, "Efficient and effective strategies for cross-corpus acoustic emotion recognition," *Neurocomputing*, vol. 275, pp. 1028–1034, 2018.

[4] T. Neuberger, A. Beke, and M. Gósy, "Acoustic analysis and automatic detection of laughter in Hungarian spontaneous speech," in *Proceedings of ISSP*, 2014, pp. 281–284.

[5] L. S. Kennedy and D. P. W. Ellis, "Laughter detection in meetings," in *Proceedings of the NIST Meeting Recognition Workshop at ICASSP*, Montreal, Canada, 2004, pp. 118–121.

[6] G. Gosztolya, A. Beke, T. Neuberger, and L. Tóth, "Laughter classification using Deep Rectifier Neural Networks with a minimal feature subset," *Archives of Acoustics*, vol. 41, no. 4, pp. 1–10, 2016.

[7] F. Grèzes, J. Richards, and A. Rosenberg, "Let me finish: Automatic conflict detection using speaker overlap," in *Proceedings of Interspeech*, Lyon, France, Aug 2013, pp. 200–204.

[8] H. Kaya, T. Özkaptan, A. A. Salah, and F. Gürgen, "Random discriminative projection based feature selection with application to conflict recognition," *IEEE Signal Processing Letters*, vol. 22, no. 6, pp. 671–675, 2015.

[9] G. Gosztolya and L. Tóth, "DNN-based feature extraction for conflict intensity estimation from speech," *IEEE Signal Processing Letters*, vol. 24, no. 12, pp. 1837–1841, 2017.

[10] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The Interspeech 2014 computational paralinguistics challenge: Cognitive & physical load," in *Proceedings of Interspeech*, Sep 2014, pp. 427–431.

[11] I. Hoffmann, D. Németh, C. Dye, M. Pákáski, T. Irinyi, and J. Kálmán, "Temporal parameters of spontaneous speech in Alzheimer's disease," *International Journal of Speech-Language Pathology*, vol. 12, no. 1, pp. 29–34, 2010.

[12] J.-R. Orozco-Arroyave, J. Arias-Londono, J. Vargas-Bonilla, and E. Nöth, "Analysis of speech from people with Parkinson's disease through nonlinear dynamics," in *Proceedings of NoLISP*, 2013, pp. 112–119.

[13] G. Kiss, M. G. Tulics, D. Sztahó, and K. Vicsi, "Language independent detection possibilities of depression by speech," in *Proceedings of NoLISP*, 2016, pp. 103–114.

[14] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[15] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The Interspeech 2015 computational paralinguistics challenge: Nativeness, Parkinson's & eating condition," in *Proceedings of Interspeech*, Dresden, Germany, Sep 2015, pp. 478–482.

[16] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The Interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proceedings of Interspeech*, San Francisco, CA, USA, Sep 2016, pp. 2001–2005.

[17] B. Schuller, S. Steidl, A. Batliner, S. Hantke, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. S. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The Interspeech 2017 computational paralinguistics challenge: Addressee, Cold & Snoring," in *Proceedings of Interspeech*, Aug 2017, pp. 3442–3446.

[18] F. Eyben, *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*. Springer, 2016.

[19] Y. Liu, W.-L. Zhao, C.-W. Ngo, C.-S. Xu, and H.-Q. Lu, "Coherent bag-of audio words model for efficient large-scale video copy detection," in *Proceedings of ACII*, Xi'an, China, Jul 2010, pp. 89–96.

[20] S. Pancoast and M. Akbacak, "Bag-of-Audio-Words approach for multimedia event classification," in *Proceedings of Interspeech*, Portland, OR, USA, Sep 2012, pp. 2105–2108.

[21] F. B. Pokorny, F. Graf, F. Pernkopf, and B. W. Schuller, "Detection of negative emotions in speech signals using bags-of-audio-words," in *Proceedings of ACII*, Xi'an, China, Sep 2015, pp. 1–5.

[22] M. Schmitt, F. Ringeval, and B. Schuller, "At the border of acoustics and linguistics: Bag-of-Audio-Words for the recognition of emotions in speech," in *Proceedings of Interspeech*, San Francisco, CA, USA, 2016, pp. 495–499.

[23] H. Lim, M. J. Kim, and H. Kim, "Robust sound event classification using LBP-HOG based Bag-of-Audio-Words feature representation," in *Proceedings of Interspeech*, Dresden, Germany, Sep 2015, pp. 3325–3329.

[24] M. Schmitt, C. Janott, V. Pandit, K. Qian, C. Heiser, W. Hemmert, and B. Schuller, "A Bag-of-Audio-Words approach for snore sounds' excitation localisation," in *Proceedings of Speech Communication*, Paderborn, Germany, Oct 2016, pp. 89–96.

[25] J. Odell, "The use of context in large vocabulary speech recognition," Ph.D. dissertation, University of Cambridge, 1995.

[26] G. Gosztolya, T. Grósz, Gy. Szaszák, and L. Tóth, "Estimating the sincerity of apologies in speech by DNN rank learning and prosodic analysis," in *Proceedings of Interspeech*, San Francisco, CA, USA, Sep 2016, pp. 2026–2030.

[27] S. Hantke, F. Weninger, R. Kurle, F. Ringeval, A. Batliner, A. E.-D. Mousa, and B. Schuller, "I hear you eat and speak: Automatic recognition of eating condition and food type, use-cases, and impact on ASR performance," *PLoS One*, pp. 1–24, 2016.

[28] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of ACM Multimedia*, 2010, pp. 1459–1462.

[29] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011.

[30] G. Gosztolya, R. Busa-Fekete, T. Grósz, and L. Tóth, "DNN-based feature extraction and classifier combination for child-directed speech, cold and snoring identification," in *Proceedings of Interspeech*, Stockholm, Sweden, Aug 2017, pp. 3522–3526.

[31] K. J. Han, S. Kim, and S. S. Narayanan, "Strategies to improve the robustness of Agglomerative Hierarchical Clustering under data source variation for speaker diarization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 8, pp. 1590–1601, 2008.

[32] W. Wang, P. Lu, and Y. Yan, "An improved hierarchical speaker clustering," *Acta Acustica*, vol. 33, no. 1, pp. 9–14, 2008.

[33] H. Kaya, A. A. Karpov, and A. A. Salah, "Fisher Vectors with cascaded normalization for paralinguistic analysis," in *Proceedings of Interspeech*, Dresden, Germany, Sep 2015, pp. 909–913.

[34] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, 2010, pp. 249–256.

[35] L. Tóth, "Phone recognition with deep sparse rectifier neural networks," in *Proceedings of ICASSP*, 2013, pp. 6985–6989.

[36] N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, "Application of pretrained deep neural networks to large vocabulary conversational speech recognition," Dept. Comp. Sci., University of Toronto, Tech. Rep., 2012.

[37] L. Tóth, "Phone recognition with hierarchical Convolutional Deep Maxout Networks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 25, pp. 1–13, 2015.

[38] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[39] T. Grósz and L. Tóth, "A comparison of Deep Neural Network training methods for Large Vocabulary Speech Recognition," in *Proceedings of TSD*, Pilsen, Czech Republic, 2013, pp. 36–43.

[40] M. Schmitt and B. Schuller, "openXBOW – introducing the Passau open-source crossmodal Bag-of-Words toolkit," *The Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2017.

[41] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.