Visual Speech for Obstructive Sleep Apnea Detection

Catarina Botelho[1], *Alberto Abad*[1], *Tanja Schultz*[2], *Isabel Trancoso*[1]

[1]INESC-ID/Instituto Superior Técnico, University of Lisbon, Portugal
[2]Cognitive Systems Lab (CSL), University of Bremen, Germany

catarina.t.botelho@tecnico.ulisboa.pt

Abstract

Obstructive sleep apnea (OSA) affects almost one billion people worldwide and limits peoples' quality of life substantially. Furthermore, it is responsible for significant morbidity and mortality associated with hypertension, cardiovascular diseases, work and traffic accidents. Thus, the early detection of OSA can save lives. In our previous work we used speech as biomarker for automatic OSA detection. More recently, we leveraged the fact that OSA patients have anatomical and functional abnormalities of the upper airway and an altered craniofacial morphology, and therefore explore information from facial images for OSA detection. In this work, we propose to combine speech and facial image information to detect OSA from YouTube vlogs. This in-the-wild data poses an inexpensive alternative to standard data collected for medical applications, which is often scarce, imbalanced and costly to acquire. Besides speech and facial images, we propose to include *visual speech* as a third modality, inspired by the emerging field of silent computational paralinguistics. We hypothesize that embeddings trained from lip reading integrate information on the craniofacial structure, on speech articulation and breathing patterns, thus containing relevant cues for OSA detection. Fusion of the three modalities achieves an accuracy of 82.5% at the speaker level.

Index Terms: Obstructive sleep apnea, silent paralinguistics, multimodality, visual speech, transfer learning, in-the-wild vlogs

1. Introduction

Obstructive sleep apnea (OSA) is a sleep-concerned breathing disorder characterized by a complete stop or decrease of the breathing airflow, despite continued or increased inspiratory efforts, while the subject is asleep [1], which leads to oxygen desaturation and sleep fragmentation [2]. This disturbed breathing impacts the quality of life of patients, who report excessive daytime sleepiness, depression, cognitive impairment [3], and mood and personality changes [4]. This disease is estimated to affect approximately one billion adults worldwide, and the number is expected to grow with population aging and obesity [5]. Besides the consequences for patients and their families, untreated OSA imposes extensive costs to the society at large, regarding economy, health system and public safety [6]. Different studies [7, 8] have estimated the economic savings that would be possible by diagnosing and treating OSA patients.

The gold standard method for OSA diagnosis, a polysomnography (PSG), cannot keep up with the growing number of cases and is thus unlikely to meet future demands [9]. Several authors have researched alternative methods for automatic OSA detection, using machine learning algorithms fed by biomarkers derived for example from facial images [10, 11, 12, 13], from speech [14, 15, 16, 17], or both [18]. The success of these biomarkers builds on the fact that OSA patients have anatomical and functional abnormalities of the upper airway and an altered craniofacial morphology [10], which impacts both some characteristics of speech and facial expressions. In our recent work [19], we focused on the altered craniofacial morphology, and therefore explored facial images for OSA detection. For that work, we collected a pilot of corpus of in-the-wild data, composed of YouTube vlogs of 40 subjects, roughly half of which claimed to suffer from OSA.

In this work we propose to use the same corpus to explore the combination of three different modalities: facial images, acoustic speech and visual speech. There is a long history of research on visual speech recognition [20, 21, 22], and a growing interest in audio-visual speech enhancement and separation [23, 24]. Nevertheless, to the best of our knowledge, visual speech has not yet been explored in the context of Silent Computational Paralinguistics, i. e., the assessment of speaker states and traits from non-audibly spoken communication [25, 26]. We argue that embeddings trained for lip reading also encode information on the craniofacial structure, speech articulation and breathing patterns. For the particular problem of OSA detection, which has been shown to benefit from queues from both speech and craniofacial structure, we hypothesize that visual speech may conjugate relevant information from both domains. Furthermore, this modality may be robust when using in-the-wild data, in which the speech signal is often contaminated with music, noise and other voices.

Although we frame this problem with in-the-wild data, which is easier to acquire in larger scale, we expect that the three modalities can be of great relevance in the context of online medical appointments or remote population screenings where both audio and visual data is available.

2. Related Work

Certain craniofacial features, including skeletal and soft tissue components, constitute a risk for the development of OSA [2]. Skeleton risk factors include a shorter mandible corpus, smaller mandibular enclosure area, retrognathia of the mandible, maxillary constriction and shorter length, narrow cranial base, inferiorly positioned hyoid bone, longer anterior face and extended head position. Soft tissue risk factors include enlarged tongue, uvula and soft palate, larger lateral pharyngeal wall and parapharyngeal fat pad volumes, smaller upper airway space and imbalance between tongue size and craniofacial enclosure [2, 6].

This motivated several researchers to use facial images and speech signals for OSA detection. These works have mostly proposed systems based on classic machine learning methods and knowledge-based features. Regarding facial images, most works use sets of craniofacial measurements derived from facial landmarks either manually or automatically labelled in facial images, including features such as mandibular length, binocular

width and face width. For speech, the most common acoustic features are Mel frequency cepstral coefficients (MFCC), linear prediction cepstral coefficients (LPCC), Energy, Harmonics-to-noise ratio (HNR), jitter, and formants frequency and bandwidth. Of particular relevance to this work is the research by Espinoza-Cuadros et al. [18] which compares speech features (i-vectors), craniofacial measurements and clinical variables for estimating OSA severity. The authors observed that i-vectors attained weaker performance than craniofacial measurements.

Fewer works have leveraged deep learning methodologies for OSA detection. An exception is the work proposed by Perero-Codosero et al. [17] which explores *x-vectors* [27] and domain adversarial training for OSA detection from speech.

In this work, we explore a third modality, inspired by the recent progress in audio visual recognition, particularly for in the wild datasets. This progress has been shown for the Lip Reading in the Wild (LRW) Dataset [21], which consists of a collection of video clips from over 1000 speakers from BBC programs, and a vocabulary of 500 English words. Each video clip is 1 second long, and contains the target word surrounded by other context words. The lip reading task is then framed as a multi-class classification problem, which predicts, for each video clip, the target word spoken. Many approaches have been proposed to tackle this challenging task, among which the ensemble of methods proposed by Ma and Martinez et al. [22]. Each method consists of a modified ResNet-18 which leverages born-again distillation (iterative self-distillation) for improving the performance. Their best single model corresponds to the third generation of knowledge distillation, and achieves a top-1 accuracy of 87.9% on the LRW dataset. The ensemble of methods achieved a state of the art top-1 accuracy of 88.5%, a promising value for the use of visual speech as an extra modality in paralinguistic tasks.

3. Methods

We address OSA automatic detection using three different modalities, facial images, acoustic speech, and visual speech. For acoustic speech and visual speech modalities, we explore transfer learning from related tasks for which large amounts of data are available and thus enable robust representation learning. Afterwards, we feed the representations to neural networks (NN), trained in a leave-one-subject-out cross-validation setting to work around the limited number of subjects. Finally, we perform a fusion of the three modalities. A schematic representation of the main steps of the best performing system for each modality is depicted in the Figure 1.

3.1. OSA detection using facial images

Data cleaning and pre-processing For each vlog, we started by extracting the key frames, several of which do not include the main subject's face, or include it with occlusions and in different angles, due to the in-the-wild nature of the data. Thus, we start by a data cleaning and pre-processing step, which includes:

- *Face detection*, using the pre-trained deep neural network face detector in *OpenCV* [28], which is based on the Single Shot Detector [29], using a ResNet-10.
- *Extraction of 68 facial landmarks*, using *Dlib*'s [30] pre-trained model, which was implemented based on [31].
- *Outlier removal*, based on interquartile range (IQR) scores. To perform this outlier removal step, each image is represented by a set of five craniofacial measurements (knowledge-based features), described below.

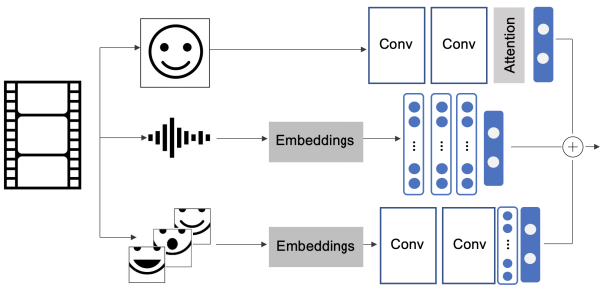

Figure 1: *Methodology pipeline.*

Feature Extraction In our recent work [19], we explored three different image representations for OSA binary classification: knowledge-based features based on the state of the art for OSA detection using facial images, bio-inspired features (originally proposed by [32]), and face embeddings trained for face recognition [33]. We compared these representations to using the raw images fed directly into an end-to-end classification system, and we achieved better results with the raw images. Hence, in this work we decided to follow that approach and use directly the raw images, without any feature extraction step.

Classification Each image, cropped around the face and resized to 100×100 was fed to a convolutional NN, which includes two convolutional blocks, one attention block, and one output linear layer with 2 nodes, followed by a softmax activation layer. Each convolutional block contains (1) one 2D-convolutional layer, with 8 filters, kernel size of 3, stride 1, and no padding; (2) one batch normalization layer; (3) one Rectified Linear Unit (ReLU) activation layer; (4) one max pooling layer with kernel size 2, and stride 2; (5) one dropout layer with dropout probability of 0.5. The attention block consists of: (1) one 2D-convolutional layer, with 8 filters, a kernel size of 3, stride 1, and padding to keep the dimension of the input constant; (2) one Sigmoid activation layer. The output of the attention block, the attention scores, are multiplied element-wise by the output of the second convolutional block, before being fed to the output layer. The convolutional layer in the attention block introduces the idea of local attention. The learning rate resembles 0.0001, the batch size was set to 32, and each network was trained for 20 epochs, using cross entropy loss.

3.2. OSA detection using acoustic speech

Data cleaning and pre-processing The original audio files were converted to mono and downsampled to 16kHz. Afterwords, we experimented using a python interface of the WebRTC Voice Activity Detector (VAD) [34], but we observed that it yielded several music, sound effects and noise segments as speech. We opted to perform speech/non-speech segmentation using a feed-forward NN trained with perceptual linear prediction (PLP) features, followed by a finite state machine, trained with broadcast news [35]. Nevertheless, the segmentation was not free or errors and some music/sound effects were still introduced as speech segments in the analysis.

Feature extraction We compare three types of features: *x-vectors*, embeddings extracted with the *problem-agnostic speech encoder*, and a *knowledge-based* feature set.

X-vectors are deep neural network based speaker embeddings, proposed as an alternative to i-vectors for speaker [27]

and language recognition [36] tasks. Besides being the current state-of-the-art for speaker recognition, *x-vectors* have also been shown to carry paralinguistic information. Several works have used them for the automatic detection of OSA [17], Parkinson's disease (PD) [37, 38], Alzheimer's disease [39], and depression [37]. Prior to *x-vector* extraction, we extract 30 MFCCs computed every 10 ms from 25 ms-length frames. We apply again a VAD to filter out remaining non-speech frames, and perform cepstral mean and variance normalization. Finally, we extract the *x-vector* embeddings with dimension 512. These steps were performed following the egs/voxceleb/v2 Kaldi recipe, and the corresponding pre-trained model [40].

The *problem agnostic speech encoder (PASE+)* [41, 42] is an encoder trained for robust representation learning in a self-supervised setting. It combines a convolutional encoder followed by multiple workers, tasked to solve self-supervised problems. PASE+ features are extracted at the end of the encoder after joint training of the encoder and the workers. These features have been shown to contain relevant information for paralinguistic tasks such as emotion recognition [41]. In order to explore their potential for automatic OSA detection, we extract PASE+ features, using the pre-trained model. Each audio segment is thus represented by a matrix with the dimension $256 \times n\_frames$.

The *knowledge-based* (KB) feature set corresponds to 109 features, proposed in [15] for OSA detection.

Classification To perform OSA binary classification using speech signals, each of the three feature sets described above was fed to a different NN. The NN architecture depends on the input type: x-vectors and KB features represent each audio segment with a fixed size vector, and thus are fed a a fully connected feed forward NN; PASE+ features represent each audio file with a dimension that depends on the segment duration, and thus are fed to a 1D convolutional network. All three NN were trained using Adam optimizer and cross entropy loss, in which each class was weighted by the inverse of it's relative frequency in the training folds.

The fully connected NNs consist of three fully connected blocks and one output linear layer of size 2, followed by a softmax activation layer. Each fully connected block consists of: (1) one linear layer with 32 nodes; (2) one batch normalization layer; (3) one ReLU activation layer; (4) one dropout layer. Both fully connected NNs were trained for 10 epochs, with batch size 64 and learning rate 0.001. The dropout probability was set to 0.5 and 0.7 in the NN trained with x-vectors and in the NN with KB features, respectively.

The convolutional NN which receives as input the PASE+ features consists of two convolutional blocks, one statistical pooling layer which summarizes the time dimension to a fixed size output, one fully connected block and one output linear layer with 2 nodes, followed by a softmax activation layer. Each convolutional block contains (1) one 1D-convolutional layer, which performs the convolution through time, with 64 and 32 filters (first and second blocks, respectively), kernel size of 3, stride 1, and padding to keep the time dimension constant; (2) one batch normalization layer; (3) one leaky ReLU activation layer; and (4) one dropout layer. The fully connected block consists of (1) one linear layer with 32 nodes; (2) one batch normalization layer; (3) one leaky ReLU activation layer; (4) one dropout layer. The learning rate resembles 0.0001, the batch size was set to 32, the dropout probability was set to 0.7, and the network trained for 10 epochs.

3.3. OSA detection using visual speech

Data cleaning and pre-processing To leverage the visual speech modality for OSA detection, we start by a pre-processing stage, which includes the following steps:

- *Voice activity detection* performed using a python interface of the WebRTC Voice Activity Detector [34].
- *One-second segmentation* of the segments for which voice was detected. Unlike the methodology used in [21], we do not attempt to perform a segmentation which includes a given target word, because we are not aiming to classify any particular words. Instead, our goal is to capture articulation and breathing patterns together with the craniofacial morphology, which can encode paralinguistic information. Thus, we segment the units resultant from VAD sequentially, into one-second segments.
- *Face detection* in all the frames of the one-second voice clips. For this step, we used the model described in section 3.1. All the clips for which the face detector did not find a face in at least 75% of the frames were excluded.
- *Extraction of 68 facial landmarks* (description in 3.1).
- *Face alignment* to a reference frame, following [22].
- *Cropping mouth region of interest* to a bounding box 96×96, following [22].

Feature extraction After the pre-processing stage we use Ma and Martinez et al. pre-trained ResNet-18 MS-TCN 3^{rd} generation student model [22] to extract lip-reading embeddings. Each clip is represented by a matrix of size $n\_frames \times 512$. Considering that not all vlogs have the same frame rate, we then used only the first 25 frames of each one-second clip.

Classification The 25×512 lip reading embeddings are fed to a convolutional NN trained for OSA's binary classification. The network contains two convolutional blocks, one fully connected block and one output linear layer with 2 nodes, followed by a softmax activation layer. Each convolutional block contains (1) one 1D-convolutional layer, which performs the convolution through time, with 64 and 32 filters (first and second blocks, respectively), kernel size of 3, stride 1, and no padding; (2) one batch normalization layer; (3) one ReLU activation layer; (4) one max pooling layer with kernel size 2, and stride 2; and (5) one dropout layer with dropout probability of 0.5. The fully connected block consists of (1) one linear layer with 32 nodes; (2) one batch normalization layer; (3) one leaky ReLU activation layer; and (4) one dropout layer.

The learning rate resembles 0.0001, the batch size was set to 64, the NN was trained for 10 epochs using cross entropy loss, and each class was weighted by the inverse of it's relative frequency in the training folds.

3.4. Fusion of the modalities

After having an OSA prediction for each subject, for each modality, we perform a majority vote to assign a final prediction for each subject. We use the predictions of the best performing system of each modality. We opted not to perform early fusion because the pre-processing of the three modalities was done independently from each other. Thus, the segments are not synchronous in time, and their quantity is different across the three modalities. Future work will tackle a synchronous pre-processing to allow early-fusion.

4. Corpus

The pilot corpus used consists of 40 vlogs publicly available at YouTube. In 22 out of the 40 vlogs, the main subjects claim to suffer from OSA, and the remaining vlogs, which serve as control subjects, were queried using unrelated keywords, such as "book review", "lets talk vlog", and "lets talk knitting". The selection of the OSA patients and controls was carefully performed in order to collect the same number of videos featuring male and female subjects, for both classes. Further details on the corpus collection can be found in [19].

While we are quite aware that the labels in this dataset are noisy and not medically verified, our approach is motivated by the success of prior work with similar datasets for the detection of PD, depression [43, 37], and OSA [15].

Although the subjects claim to suffer from OSA and talk about their disease, we believe that it does not compromise the results using lip reading embeddings. We manually verified that the vocabulary of the LRW, used for training the lip reading extractor, does not contain the target words "obstructive sleep apnea", "apnea", "sleep", "disease" nor "disorder". The only words in the vocabulary related to healthcare were "hospital" and "medical". For the vlogs for which the automatic transcription was publicly available (38 out of 40) we counted the number of occurrences of those two words. The word hospital occurs 5 times (0.02%), all of them spoken by subjects with OSA, and the word "medical" occurs 35 times (0.16%), 32 of which spoken by subjects with OSA.

The different pre-processing steps for each modality described in section 3 result in different numbers of instances per modality. The summary of the number of instances per speaker, per modality is presented in Table 1. The large standard deviations indicate a large variation in the count of instances per subject – while some subjects may have very few instances others have several hundreds. All subjects have at least 5 instances in each modality.

Table 1: *Dataset description: instance counts per modality.*

Modality	Total count	count per subject mean ± std
Facial images	2733	68 ± 56
Audio files	4953	124 ± 133
Video clips	22261	557 ± 401

5. Results

Table 2 presents the classification results for the different modalities and the fusion results. Figure 2 shows the majority vote predictions attributed to each speaker. Considering that the training mode was leave-one-subject-out cross validation setting, we present three accuracy metrics: *accuracy* is computed with the ratio of correctly classified instances on all cross-validation folds, over all instances of the dataset; *mean accuracy per subject* results from computing the mean of the accuracies obtained in each fold (i.e. each subject); and *majority vote accuracy* results of first averaging the predictions of all predictions of each subject, obtaining one single prediction for each subject, and then computing the accuracy. This last metric leverages the fact that we have multiple instances for each subject, and compensates for the fact that some instances maybe noisy or even contain different subjects.

To facilitate result interpretation, Table 2 also shows the chance level, or *prior*, for each modality, which corresponds to having a classifier that always predicts the most frequent class in the dataset. The prior for accuracy differs across modalities because the number of instances also differs; the prior for subject-level accuracy metrics remains constant because we are assuming a classifier based solely on the number of subjects, which remains constant for all modalities.

Table 2: *Accuracy results (%) of the classification experiments.*

Modality		Acc	Mean acc per subject	Majority vote acc
Facial images	prior	53.5	55.0	55.0
	results	**76.3**	**73.4**	77.5
Speech	prior	50.8	55.0	55.0
	x-vectors	65.1	62.9	67.5
	PASE+	60.3	62.3	62.5
	KB	55.6	55.5	55.0
Visual speech	prior	51.7	55.0	55.0
	results	69.8	69.6	**80.0**
Fusion		–	–	**82.5**

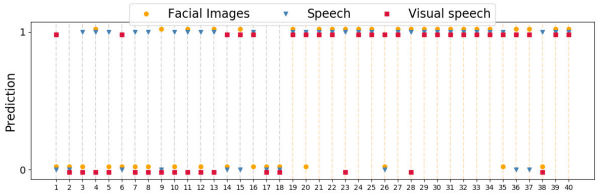

Figure 2: *Rounded predictions per subject, using each modality. Subjects 1 to 18 are the controls, and 19 to 40 are OSA patients.*

Overall, we observe that the facial images modality achieves the best results at the instance level, and visual speech achieves the best results after performing a majority vote. The speech modality performs worse than the other two. We hypothesize that this may be due to the fact that several audio segments are contaminated with music and noise. The use of a VAD module that had been designed for broadcast news may not be robust enough for the vlog domain and must be replaced in future work. Comparing the three speech representations, we observe that x-vectors outperform the other two representations.

6. Conclusions

This work compares and merges three different modalities, facial images, speech, and visual speech for OSA automatic detection, using transfer learning. This work builds on tools that allowed us to explore the contribution of each modality separately. Nevertheless, we believe that a careful alignment of the input segments for each modality would enable early fusion and potentially leverage greater synergies between the three modalities. To the best of our knowledge, this is the first work that explores visual speech for paralinguistic applications. The results obtained are promising, and we anticipate that can be further improved with data from more controlled conditions such as the case of remote medical appointments.

7. Acknowledgements

This work was supported by national funds through FCT, Fundação para a Ciência e a Tecnologia, under project UIDB/50021/2020 and grant number SFRH/BD/149126/2019.

8. References

[1] J. Arnold, M. Sunilkumar, V. Krishna, S. Yoganand, M. S. Kumar, and D. Shanmugapriyan, "Obstructive sleep apnea," *Journal of pharmacy & bioallied sciences*, vol. 9, no. Suppl 1, p. S26, 2017.

[2] K. Sutherland, R. W. Lee, and P. A. Cistulli, "Obesity and craniofacial structure as risk factors for obstructive sleep apnoea: impact of ethnicity," *Respirology*, vol. 17, no. 2, pp. 213–222, 2012.

[3] N. Punjabi, "The epidemiology of adult obstructive sleep apnea," *Proceedings of the American Thoracic Society*, vol. 5, no. 2, pp. 136–143, 2008.

[4] T. Paiva, M. Andersen, and S. Tufik, "Sono e a medicina do sono. 1ª edição," 2014.

[5] A. V. Benjafield, N. T. Ayas, P. R. Eastwood, R. Heinzer, M. S. Ip, M. J. Morrell, C. M. Nunez, S. R. Patel, T. Penzel, J.-L. Pépin *et al.*, "Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis," *The Lancet Respiratory Medicine*, vol. 7, no. 8, pp. 687–698, 2019.

[6] M. M. Lyons, N. Y. Bhatt, A. I. Pack, and U. J. Magalang, "Global burden of sleep-disordered breathing and its implications," *Respirology*, vol. 25, no. 7, pp. 690–702, 2020.

[7] F. Sullivan, "Hidden health crisis costing america billions: Underdiagnosing and undertreating obstructive sleep apnea draining healthcare system," *American Academy of Sleep Medicine*, 2016.

[8] P. Armeni, L. Borsoi, G. Donin, F. Costa, and L. Ferini-Strambi, "Pnd33 the clinical and economic burden of obstructive sleep apnea in adults: A cost-of-illness analysis," *Value in Health*, vol. 22, p. S742, 2019.

[9] P. de Chazal, P. A. Cistulli, and M. T. Naughton, "The future of sleep-disordered breathing: A public health crisis," *Respirology*, 2020.

[10] A. Balaei, K. Sutherland, and P. Cistulli et al., "Automatic detection of obstructive sleep apnea using facial images," in *ISBI*. IEEE, 2017.

[11] R. W. Lee, A. S. Chan, R. R. Grunstein, and P. A. Cistulli, "Craniofacial phenotyping in obstructive sleep apnea — a novel quantitative photographic approach," *Sleep*, vol. 32, no. 1, pp. 37–45, 2009.

[12] H. Nosrati, N. Sadr, and P. de Chazal, "Apnoea-hypopnoea index estimation using craniofacial photographic measurements," in *CinC*. IEEE, 2016.

[13] A. T. Balaei, K. Sutherland, P. Cistulli, and P. de Chazal, "Prediction of obstructive sleep apnea using facial landmarks," *Physiological measurement*, vol. 39, no. 9, p. 094004, 2018.

[14] A. M. Benavides, R. F. Pozo, D. T. Toledano, J. L. B. Murillo, E. L. Gonzalo, and L. H. Gómez, "Analysis of voice features related to obstructive sleep apnoea and their application in diagnosis support," *Computer Speech & Language*, vol. 28, no. 2, pp. 434–452, 2014.

[15] C. Botelho, I. Trancoso, A. Abad, and T. Paiva, "Speech as a biomarker for obstructive sleep apnea detection," in *ICASSP*. IEEE, 2019, pp. 5851–5855.

[16] M. Kriboy, A. Tarasiuk, and Y. Zigel, "Detection of obstructive sleep apnea in awake subjects by exploiting body posture effects on the speech signal," in *EMBC*. IEEE, 2014.

[17] J. M. Perero-Codosero, F. Espinoza-Cuadros, J. Antón-Martín, M. A. Barbero-Álvarez, and L. A. Hernández-Gómez, "Modeling obstructive sleep apnea voices using deep neural network embeddings and domain-adversarial training," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 240–250, 2019.

[18] F. Espinoza-Cuadros, R. Fernández-Pozo, D. T. Toledano, J. D. Alcázar-Ramírez, E. López-Gonzalo, and L. A. Hernández-Gómez, "Speech signal and facial image processing for obstructive sleep apnea assessment," *Computational and mathematical methods in medicine*, vol. 2015, 2015.

[19] C. Botelho, T. Schultz, A. Abad, and I. Trancoso, "Binary classification of obstructive sleep apnea from facial images in-the-wild," *submitted to Computer Vision and Image Understanding*, 2021.

[20] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, "A review of recent advances in visual speech decoding," *Image and vision computing*, vol. 32, no. 9, pp. 590–605, 2014.

[21] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 87–103.

[22] P. Ma, B. Martinez, S. Petridis, and M. Pantic, "Towards practical lipreading with distilled and efficient models," in *arXiv preprint arXiv:2007.06504, accepted to ICASSP*. IEEE, 2021.

[23] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," *arXiv preprint arXiv:1711.08789*, 2017.

[24] K. Tan, Y. Xu, S.-X. Zhang, M. Yu, and D. Yu, "Audio-visual speech separation and dereverberation with a two-stage multimodal network," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 542–553, 2020.

[25] L. Diener, S. Amiriparian, C. Botelho, K. Scheck, D. Küster, I. Trancoso, B. W. Schuller, and T. Schultz, "Towards silent paralinguistics: Deriving speaking mode and speaker ID from electromyographic signals," in *Interspeech*, 2020.

[26] C. Botelho, L. Diener, D. Küster, K. Scheck, S. Amiriparian, B. W. Schuller, T. Schultz, A. Abad, and I. Trancoso, "Toward silent paralinguistics: Speech-to-emg – retrieving articulatory muscle activity from speech," in *Interspeech*, 2020.

[27] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, April 2018, pp. 5329–5333.

[28] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[30] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[31] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *CVPR*, 2014, pp. 1867–1874.

[32] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," in *CVPR*. IEEE, 2009, pp. 112–119.

[33] D. King, "High quality face recognition with deep metric learning," 2017. [Online]. Available: http://blog.dlib.net/2017/02/high-quality-face-recognition-with-deep.html

[34] J. Wiseman, "py-webrtcvad," retrieved in March 2021. [Online]. Available: https://github.com/wiseman/py-webrtcvad

[35] H. Meinedo and J. Neto, "A stream-based audio segmentation, classification and clustering pre-processing system for broadcast news using ann models," in *Interspeech*, 2005.

[36] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, 2017, pp. 999–1003.

[37] J. Correia, F. Teixeira, C. Botelho, I. Trancoso, and B. Raj, "The in-the-wild speech medical corpus," in *ICASSP*. IEEE, 2021.

[38] L. Moro-Velazquez, J. Villalba, and N. Dehak, "Using x-vectors to automatically detect parkinson's disease from speech," in *ICASSP*. IEEE, 2020, pp. 1155–1159.

[39] S. Zargarbashi and B. Babaali, "A multi-modal feature embedding approach to diagnose alzheimer disease from spoken language," *arXiv preprint arXiv:1910.00330*, 2019.

[40] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *ASRU*, Dec. 2011.

[41] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, "Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks," in *INTERSPEECH*, 2019.

[42] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-task self-supervised learning for Robust Speech Recognition," *preprint ArXiv:2001.09239*, 2020.

[43] J. Correia, B. Raj, I. Trancoso, and F. Teixeira, "Mining multimodal repositories for speech affecting diseases," in *INTERSPEECH*, 2018.