INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Automatic Assessment of Dysarthric Severity Level Using Audio-Video Cross-Modal Approach in Deep Learning

Han Tong[1], *Hamid Sharifzadeh*[1], *Ian McLoughlin*[2]

[1]School of Computing, Unitec Institute of Technology, Auckland, New Zealand
[2]ICT Cluster, Singapore Institute of Technology, Singapore
{tongh02, hsharifzadeh}@unitec.ac.nz, Ian.McLoughlin@singaporetech.edu.sg

Abstract

Dysarthria is a speech disorder that can significantly impact a person's daily life, and yet may be amenable to therapy. To automatically detect and classify dysarthria, researchers have proposed various computational approaches ranging from traditional speech processing methods focusing on speech rate, intelligibility, intonation, etc. to more advanced machine learning techniques. Recently developed machine learning systems rely on audio features for classification; however, research in other fields has shown that audio-video cross-modal frameworks can improve classification accuracy while simultaneously reducing the amount of training data required compared to uni-modal systems (i.e. audio- or video-only).

In this paper, we propose an audio-video cross-modal deep learning framework that takes both audio and video data as input to classify dysarthria severity levels. Our novel cross-modal framework achieves over 99% test accuracy on the UASPEECH dataset – significantly outperforming current uni-modal systems that utilise audio data alone. More importantly, it is able to accelerate training time while improving accuracy, and to do so with reduced training data requirements.

Index Terms: Dysarthria, CNN, Cross-modal, UASPEECH

1. Introduction

As speech generation involves a collaboration of muscles, neurones and brain areas, any damage or disease to those organs may cause problems with speech phonation. Physical injury to or impairment of the central nervous cells or peripheral nervous cells in the brain can lead to a range of speech disorders that are collectively called dysarthria [1]. People with dysarthria have difficulty controlling their lips, tongue or other speech articulators, leading to changes in speaking rate, intonation, articulation, pitch or amplitude dynamics compared to normal speech. As a result, dysarthric speech tends to have reduced intelligibility, creating obstacles for patients' daily communication, and affecting quality of life. Based on the locus and the symptoms of the disease, dysarthria is medically categorised into different types such as spastic, flaccid, ataxic, hyperkinetic, hypokinetic, and mixed [2]. To evaluate a patient's progression in the root cause of the disease, doctors need to know the severity level of the dysarthria. Depending on the locus and severity level, clinical decisions are made regarding treatment options, efficacy of therapy, or of speech recovery sessions.

Clinical assessment of dysarthria is mainly auditory based using subjective tests [3]. Classic subjective tests require the presence of a Speech-Language Pathologist (SLP), which is costly and time-consuming. Also, the subjective nature of these procedures leads to variations in the reliability and validity of the diagnosis. With the help of computational advances, practical speech processing principles, and some recently developed machine learning algorithms, much research effort has been made to find more effective and more reliable ways to accomplish such a diagnostic task with higher reliability, accuracy, and consistency, while being cheaper and faster [4, 5].

Current computational research in this context can be categorised into two broad classes of speech processing-based or machine learning-based methods. In both, only one form of data is used as a sole resource (i.e. audio data) for automatic diagnostic and feature extraction/analysis. At the time of writing, this is the first paper to introduce a cross-modal audio-video classifier of dysarthria severity level. Furthermore, our proposed classifier is one of the first to apply deep-learning through Convolutional Neural Networks (CNN) to this task.

As the audio samples of dysarthric subjects for training a classifier are usually restricted (especially compared with most related speech classification problems in which training samples are plentiful), our proposed cross-modal system aims to make the best use of the available subjects by combining audio and video data collected from each of them.

The remainder of this paper is organised as follows: Section 2 briefly reviews related work. Section 3 introduces our proposed method including audio and video data pre-processing, network architecture, feature extraction and combination. Section 4 evaluates the method while Section 5 concludes the paper.

2. Background & Related Work

In general, early intervention in the progress of a disease leads to better outcomes; for dysarthria this may mean commencing appropriate therapy as soon as possible. Automatic dysarthria assessment, which can assist healthcare professionals to diagnose faster, is therefore a potentially valuable technology.

Early computational systems which addressed dysarthria related issues used digitalised toolkits such as pathological change detection in speech caused by dysarthria [6], or digit recognition systems for dysarthric speech [7], and dysarthria symptom detection [8]. Most methods analysed dysarthric speech features such as intonation, hoarseness, intelligibility, pitch variation, speech rate, etc. while still relying on human intervention and expert knowledge to interpret the results.

However rapid advancements in machine learning in recent years led researchers to deploy training-based approaches to implement fully automatic systems for processing and diagnosis of dysarthric speech. Acoustic features from Automatic Speech Recognition (ASR), such as Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP) and Mel-Frequency Cepstral Coefficients (MFCCs), were also used in these machine learning systems for feature extraction and training [9, 10, 11]. In addition to local acoustic features, some global statistics fea-

http://dx.doi.org/10.21437/Interspeech.2020-1997

tures such as speech variance, skewness, and kurtosis were commonly used.

Nakashika et al. [4] proposed the CNN-based Convolution Bottleneck Network (CBN), to extract dysarthric speech features. Instead of using traditional acoustic features for classification, they inserted various raw features into the pre-trained CBN and extracted features from its bottleneck layer, achieving high word accuracy for dysarthric speech. A similar approach [12] employed a deep bottleneck extractor with DNN classifier (DBN-DNN) to detect disfluency in dysarthric speech. With raw MFCC and linear predictive cepstral coefficient (LPCCs) input features to the DBN, the DNN achieved high accuracy for stuttering detection in dysarthric speech.

In addition to audio classification, deep learning methods can also effectively handle video data, having been successfully applied in many areas of computer vision including face recognition, action recognition, emotion detection and lip reading [13, 14, 15, 16, 17]. Furthermore, recent research shows that by combining audio and video data (i.e. dual modality systems) the amount of required training data can be reduced while improving performance; in fact, leveraging on this advantage, cross-modal systems have already been deployed in fields such as computer vision, audio-video recognition tasks, and music video retrieval [18, 19, 20].

Despite such efforts, detection of dysarthric speech using video data alone has been absent from the literature, as have cross-modal systems. Given the limited availability of dysarthric speech data and the fact that what little is available is often accompanied by video, we propose an audio-video cross modal deep learning framework for automatic assessment of dysarthria severity levels .

3. Proposed Method

The proposed cross-modal framework for automatic assessment of dysarthria uses both audio and video data as input. Acoustic and visual features will be used jointly for dysarthric speech feature learning. The UASPEECH [21] dataset is employed in our research for training and assessment. This dataset focuses only on dysarthric speech and contains both audio and video recordings of patients with dysarthria.

3.1. Network Diagram

Figure 1 shows the overall data flow of the proposed framework in which audio and video files traverse different paths. After pre-processing, MFCCs are calculated and formed into an acoustic feature cube by front-end CNN layers. A visual feature cube is similarly formed after feature extraction and CNN processing of video data. The two cubes are then combined to form a set of joint audio-video feature vectors. Finally, through a supervised learning approach, a fully-connected neural network classifier is trained to detect the severity levels of dysarthria patients based on these joint vectors.

The CNN convolution layers are defined by

$$Z^{l+1}(i,j) = \sum_{k=1}^{K_l}\sum_{x=1}^{f}\sum_{y=1}^{f}\left[Z_k^l\left(s_0 i + x, s_0 j + y\right) w_k^{l+1}(x,y)\right] + b \tag{1}$$

where b is the bias vector, Z^l and Z^{l+1} represent the input and output of the convolution layer feature maps. $Z^{l+1}(i,j)$ corresponds to the pixels in the feature map. k is the channel number. f, s_0, and w are the parameters of the convolution layer, namely, the size of the convolution kernel, the stride of the convolution, and the amount of padding, respectively. The inputs and outputs of nodes in the neural networks can be expressed as:

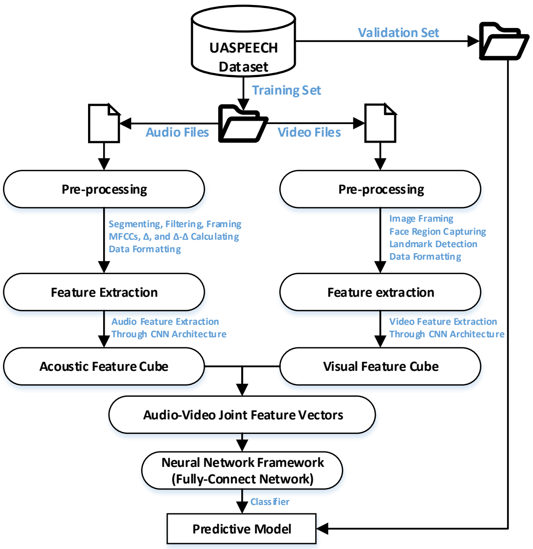

Figure 1: *Overall dataflow of the proposed cross-modal architecture.*

$$A_{i,j,k}^l = f\left(Z_{i,j,k}^l\right) \tag{2}$$

where k is the channel number in the feature map and l is the layer number. For activation functions, we use Rectified Linear Units (ReLU) functions in all layers, $f(x) = max(0,x)$ for input x.

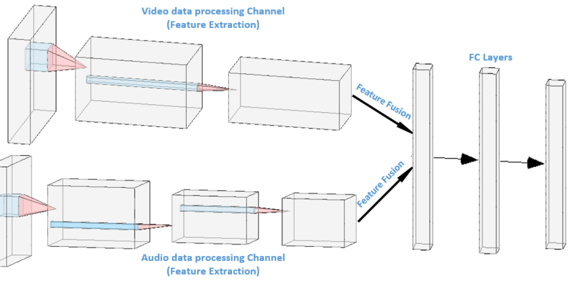

Figure 2: *Proposed audio-video cross-modal network structure.*

3.2. Cross-Modal Network Structure

Figure 2 illustrates the structure of the proposed audio-video cross-modal network. Audio and video data flow through two separate input arms which extract high level acoustic and visual features respectively. After the feature extraction stage, data fusion is applied to combine the acoustic and visual features to form joint feature vectors. Then, the combined feature vector is fed into a set of FC layers to perform joint learning and finally generate a prediction vector for the final classifier layer. A Softmax layer is used to generate 4 dimensional one-hot coded vectors as the final result. Further details on audio and video processing are discussed in Subsections 3.4 and 3.5.

Table 1: *Speaker intelligibility scores, dysarthria diagnosis and severity levels for each speaker in the dataset.*

Speaker	Age	Speech Intelligibility	Diagnosis	Severity Level
M01	>18	very low (15%)	Spastic	Severe
M04	>18	very low (2%)	Spastic	Severe
M05	21	mid (58%)	Spastic	Mild
M07	58	low (28%)	Spastic	Moderate
M08	28	high (93%)	Spastic	Low
M09	18	high (86%)	Spastic	Low
M10	21	high (93%)	Mixed	Low
M11	48	mid (62%)	Athetoid	Mild
M12	19	very low (7.4 %)	Mixed	Severe
M14	40	high (90.4%)	Spastic	Low
M16	-	low (43%)	Spastic	Moderate
F02	30	low (29%)	Spastic	Moderate
F03	51	very low (6%)	Spastic	Severe
F04	18	mid (62%)	Athetoid	Mild
F05	22	high (95%)	Spastic	Low

3.3. Dataset

UASPEECH is a dataset of dysarthric speech for research purposes containing isolated-word recordings of speakers with spastic dysarthria. Speech utterances were collected in a lab environment with 8 microphones and a digital camera. Subjects read isolated words from a computer monitor and a total of 15 speakers (4 females and 11 males) are included in the experiment. A summary table of the speakers and their relevant condition is shown in Table 1. M and F in the speaker codes indicate the gender of the speakers and is followed by a numerical identifier. M02, M03, M06, and F01 were excluded because those subjects were either recorded under a different protocol or did not approve redistribution of their data.

For each speaker, 765 words were recorded in 3 blocks of 255. Each block contains 155 common words that are repeated, plus 100 uncommon words. The repeated common words include the 10 digits and 26 letters of the alphabet, along with 19 computer command words and 100 typical common words.

The speech intelligibility score is based on average scores given in listening tests by 5 native speakers. Their ratings range widely from 2% to 95%. Aligned with the literature, we have classified the dysarthric speakers into four groups based on their speech intelligibility score, i.e. very low for 0-25%, low for 25-50%, medium for 50-75%, and high for 75-100%. The corresponding dysarthria levels are: ‘severe’, ‘moderate’, ‘mild’, and ‘low’. Table 1 shows the detailed intelligibility scores for each speaker. Based on those scores, the dysarthria severity level will be classified into the four pre-defined severity levels. For example, M12 has a very low intelligibility score (7.4%), so a correctly classified dysarthria level for M12 would be ‘severe’.

3.4. Audio Processing

The duration of the pre-segmented audio files, each of which contains one word, ranges from 1 to 20 seconds. However most sound files contain long periods of silence, especially before and after the utterance. Energy-based voice activity detection (VAD) is therefore applied to filter out unwanted information.

MFCC features are used as acoustic features. Analysis windows are 40ms long and advance with a 25ms step. The native sampling rate is 16kHz, giving 640 samples per window and 400 samples per step. In addition to raw MFCCs, we follow many ASR systems and also compute the delta and delta-delta of the MFCCs [22]. For example, for an audio window with 15 frames and 30 filter banks within each frame, 45 (15 × 3) 1-D feature maps will be generated with each map having 30 dimensions.

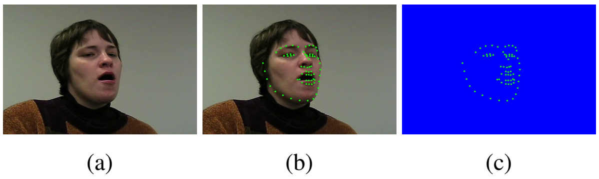

Figure 3: *a One video frame of speaker F02. b F02 facial landmarks. c Key expression information after background filtering.*

3.5. Video Processing

The video was recorded at 29.969 fps. Similar to audio data, video recordings are also segmented. Each segment contains only 18 frames. With the help of a pre-trained face detector, the face region is identified and selected as a region of interest (ROI) for every video frame. Then, facial muscle movements can be modelled through the Facial Action Coding System (FACS) [23]. To track facial muscle movement, we use facial landmarks. With the help of the dlib library, we were able to draw facial landmarks for all captured faces. A total of 68 landmarks were used to outline the shape of a face and obtain the facial features.

Figure 3 shows a frame example which was obtained from video recording block 1 of speaker F02. Figure 3a shows the original frame captured from the video. Figure 3b overlays the face detection and 68 landmark features. Figure 3c then shows only the landmarks after background filtering. The same process is applied to all the frames across all videos to track the facial movements for all speakers.

Once the position of the facial landmarks are obtained, the facial movement parameters are calculated frame by frame. More specifically, the position and angle of the face are calculated to track its overall movement, while the shape of the eyes, eyebrows and mouth are calculated to track the details of facial movement. In total, 54 measurements are calculated to track the facial movements. In fact, we define two regions as our primary feature regions; the mouth and the eyes, since they are important for emotion reading and language generation (i.e. they are also important detection points for human listeners). Specifically, these regions are the upper and lower lips, eyes, and eyebrows. Thus, for each video segment, a vector with dimensions of 18 by 54 is generated.

3.6. Combined Features

The MFCC features are delivered to a 2D-CNN network and the facial landmarks similarly delivered to a separate CNN network for feature learning. The CNN-processed audio feature vectors and video feature vectors are then combined using a late fusion technique as in [24, 25]. We used the same approach to form joint high-level acoustic-visual feature vectors that contain both video and audio information from dysarthric speakers. A flattening layer is attached to the end of the feature extraction network to flatten the extracted features.

A fully connected (FC) classifier is then trained from these joint feature vectors. As shown in Figure 2, the layers on the left are the input which takes the raw features, while the output layer on the right generates a 4 dimensional one-hot coded classification to indicate the severity level of the dysarthric input speech.

4. Results and Discussion

We trained three different types of predictive model using three different training input setups: (1) audio-only predictive models that use only the audio features as input, (2) video-only predictive models that use only video features as input, and (3) audio-video cross-modals using both audio and video data. For each setup, we trained two different network structures to gain an idea of how different architectures might influence the performance of the proposed deep-learning model. We present two matching architectures for each type of model.

As for the testing, audio and video recordings from subjects F02 (moderate), M10 (low), M11 (mild), M14 (low), and block 3 of M12 (severe) were manually selected for testing (at least one subject from each severity level group); the samples of these subjects were excluded from the network training/validation process (i.e. they were only used to test the network for reporting the accuracy results).

Each architecture we tested is defined by the number of layers in the CNN feature extractor and the number of filters per layer, as well as the number of layers and number of nodes in the back-end FC classifier. For example, the first audio-only predictive model we tested has the architecture 64-32-/-64-32-4 which means that the network has two convolutional and max pooling layers in the feature extraction CNN and three FC layers in the classifier. The CNN parameters represent the number of filters in each layer, whereas the FC parameters indicate the number of nodes in the FC layers. From this point on, we will exclude the final layer for simplicity (because it obviously has 4 nodes in each system). The audio-only networks were trained over 20 epochs, and the video-only networks over 30 epochs.

As mentioned above, the cross-modal network has two separated processing channels, one for audio data feature extraction and one for video data feature extraction. After both acoustic and visual features are obtained, they are combined to form joint features. Both networks were trained for 25 epochs.

Table 2 summarises the results for the different experimental systems and architectures. The system initial indicates the raw feature type (audio, video or both) and the suffix number indicates the number of layers in the CNN feature extractor. From these results, we can see some interesting trends emerge. Firstly, we note the affinity of audio features for the less complex network (around 0.4% higher testing accuracy for A2 vs. A3) but the converse being true for video features (around 2.8% higher testing accuracy for the more complex V4 vs. V3). In fact video features alone, in the more complex V4 system, achieved a very high testing accuracy.

Moving to the cross-model architectures, of which two kinds of feature extraction are tested (both audio and video are processed with a simpler CNN, or both are processed with a more complex CNN), we see that the more complex feature extraction system AV4 performed best, achieving a 99.5% testing accuracy. However the simpler AV3 architecture exceeded the performance of all single feature A and V systems.

The results represent the mean accuracy over all subjects. We can also examine per-subject testing accuracies in the boxplot of Figure 4 for all the predictive models identified in Table 2. From the plot, we can see that for the single modality systems there is quite a wide range of accuracies. This indicates that the method may well be quite suitable for some subjects (e.g. up to 96% even for AV2) but not for others. The range is also large for the simpler systems, indicating that the generality of the more complex systems has improved (even when, in the case of A3, this degrades the mean result). However for cross-modality systems, not only is mean accuracy much higher, but the spread of results is smaller, indicating that the system works for all users, not just some users – consequently also lending confidence in the results.

Table 2: *Summary table of experimental results*

Input data	System extraction	Classif-ication	Epochs	Train. acc. %	Test acc. %
A2	64-32	64-32	20	91.6	93.0
A3	128-64-32	128-64-32	20	99.6	92.6
V3	128-64-32	64-32	30	99.8	96.1
V4	256-128-64-32	128-64-32	30	99.5	98.9
AV3	A=128-64-32 V=128-64-32	128-64-32	25	99.8	99.1
AV4	A=256-128-64-32 V=256-128-64-32	128-64-32	25	99.8	99.5

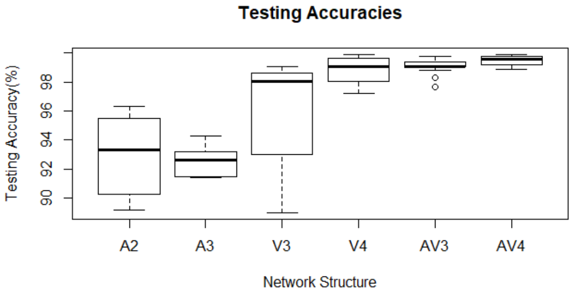

Figure 4: *Box plots summarising testing accuracies of two network structures for all models*

The proposed method also surpasses the performance of existing published classifiers operating on the same task. For example, the best performing ANN architecture used in [26] computed 31 audio descriptors of 11 feature sets to achieve an accuracy of 96.44% (around 3% less than the proposed method). Also, since they had 8 different feature vector lengths in that system, they needed to create separate ANN configurations for each of the feature sets. This complicated the entire training process. By contrast, the proposed network architecture processes the training data uniformly so there is no need to simultaneously train multiple network structures – making the overall training process much easier to configure.

5. Conclusion

This paper has proposed a novel cross-modal method for automatic assessment of dysarthria. For the first time, we have combined audio and video features and jointly used them for classifier learning. We have also validated the use of facial landmarks for predicting dysarthria severity. Our final cross-modal deep-learning framework provides a relatively low complexity method for dysarthria severity level assessment that, by leveraging the complementary strengths of both audio and video data, can improve on single modality classification. In fact, it achieves 99% test accuracy on the UASPEECH dataset, comfortably outperforming the current state-of-the-art.

In future we aim to assess alternative architectural designs as well as different audio feature representations. Our eventual aim is to produce technology which can assist physicians in early diagnosis of dysarthria severity, enabling them to admit patients to therapy sessions earlier, and more effectively track recovery progress.

6. References

[1] P. Enderby, "Disorders of communication: Dysarthria," *Handbook of clinical neurology / edited by P.J. Vinken and G.W. Bruyn*, vol. 110, pp. 273–81, 01 2013.

[2] B. J. Zyski and B. E. Weisiger, "Identification of dysarthria types based on perceptual analysis," *Journal of Communication Disorders*, vol. 20, no. 5, pp. 367–378, Oct. 1987.

[3] T. Schölderle, E. Haas, and W. Ziegler, "Age norms for auditory-perceptual neurophonetic parameters: A prerequisite for the assessment of childhood dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 63, pp. 1–12, 04 2020.

[4] T. Nakashika, T. Yoshioka, T. Takiguchi, Y. Ariki, S. Duffner, and C. Garcia, "Dysarthric speech recognition using a convolutive bottleneck network," in *2014 12th International Conference on Signal Processing (ICSP)*, Oct. 2014, pp. 505–509.

[5] J. N. Carmichael, "Enhancing speech rate estimation techniques to improve dysarthria diagnosis," in *2017 8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, Oct 2017, pp. 309–313.

[6] T. Orzechowski, A. Izworski, R. Tadeusiewicz, K. Chmurzynska, P. Radkowski, and I. Gatkowska, "Processing of pathological changes in speech caused by dysarthria," in *2005 International Symposium on Intelligent Signal Processing and Communication Systems*, Dec 2005, pp. 49–52.

[7] M. Hasegawa-Johnson, J. Gunderson, A. Perlman, and T. Huang, "Hmm-based and svm-based recognition of the speech of talkers with spastic dysarthria," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings (ICASSP)*, May 2006, pp. III–III.

[8] J. Carmichael, "Dysarthria diagnosis via respiration and phonation," in *2015 International Conference and Workshop on Computing and Communication (IEMCON)*, Oct. 2015, pp. 1–5.

[9] G. Vyas, M. K. Dutta, J. Prinosil, and P. Harár, "An automatic diagnosis and assessment of dysarthric speech using speech disorder specific prosodic features," in *2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, Jun. 2016, pp. 515–518.

[10] T. B. Ijitona, J. J. Soraghan, A. Lowit, G. Di-Caterina, and H. Yue, "Automatic detection of speech disorder in dysarthria using extended speech feature extraction and neural networks classification," in *IET 3rd International Conference on Intelligent Signal Processing (ISP 2017)*, Dec. 2017, pp. 1–6.

[11] A. Benba, A. Jilbab, A. Hammouch, and S. Sandabad, "Voiceprints analysis using MFCC and SVM for detecting patients with Parkinson's disease," in *2015 International Conference on Electrical and Information Technologies (ICEIT)*, Mar. 2015, pp. 300–304.

[12] S. Oue, R. Marxer, and F. Rudzicz, "Automatic dysfluency detection in dysarthric speech using deep belief networks," in *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, Sep. 2015, pp. 60–64.

[13] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based Emotion Recognition Using CNN-RNN and C3d Hybrid Networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI)*, 2016, pp. 445–450.

[14] I. Fung and B. Mak, "End-To-End Low-Resource Lip-Reading with Maxout Cnn and Lstm," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 2511–2515.

[15] Z. Xu, V. Vilaplana, and J. R. Morros, "Action Tube Extraction Based 3d-CNN for RGB-D Action Recognition," in *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, Sep. 2018, pp. 1–6.

[16] X. Chen and Y. Han, "Multi-task CNN Model for Action Detection," in *2018 IEEE Visual Communications and Image Processing (VCIP)*, Dec. 2018, pp. 1–4.

[17] W. Lan, J. Dang, Y. Wang, and S. Wang, "Pedestrian detection based on yolo network model," in *2018 IEEE International Conference on Mechatronics and Automation (ICMA)*, Aug 2018, pp. 1547–1551.

[18] A. Brutti and A. Cavallaro, "Online Cross-Modal Adaptation for Audio–Visual Person Identification With Wearable Cameras," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 1, pp. 40–51, Feb. 2017.

[19] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, and J. Dawson, "3d Convolutional Neural Networks for Cross Audio-Visual Matching Recognition," *IEEE Access*, vol. 5, pp. 22 081–22 091, 2017.

[20] D. Zeng, Y. Yu, and K. Oyama, "Audio-Visual Embedding for Cross-Modal Music Video Retrieval through Supervised Deep CCA," in *2018 IEEE International Symposium on Multimedia (ISM)*, Dec. 2018, pp. 143–150.

[21] H. Kim, M. A. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Dec. 2008, pp. 1741–1744.

[22] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, Oct 2014.

[23] T. R. Brick, M. D. Hunter, and J. F. Cohn, "Get the facs fast: Automated facs face analysis benefits from the addition of velocity," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, Sep. 2009, pp. 1–7.

[24] Y. Dong, S. Gao, K. Tao, J. Liu, and H. Wang, "Performance evaluation of early and late fusion methods for generic semantics indexing," *Pattern Analysis and Applications*, vol. 17, no. 1, pp. 37–50, Feb. 2014.

[25] M. Seeland, M. Rzanny, N. Alaqraa, J. Wäldchen, and P. Mäder, "Plant species classification using flower images—a comparative study of local feature representations," *PLOS ONE*, vol. 12, p. e0170629, Feb. 2017.

[26] C. Bhat, B. Vachhani, and S. K. Kopparapu, "Automatic assessment of dysarthria severity level using audio descriptors," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 5070–5074.