Digital Signal Processing and System Theory
Prof. Dr.-Ing. Gerhard Schmidt

A Review of Audio-visual Speech Recognition Approaches

Seminar Report for
'Selected Topics in Digital Signal Processing'

written by
B. Sc. Abidur Rahman

supervised by
Prof. Dr.-Ing. Gerhard Schmidt (first supervisor)
M. Sc. Karolin Krüger (second supervisor)

February 02, 2024

Declaration

I declare that I have produced the seminar report

A Review of Audio-visual Speech Recognition Approaches

independently and without improper external assistance and that I have identified all word-for-word quotations of other authors, as well as comments based closely on other authors' ideas, and I have listed the relevant sources.

I am aware that, unless agreed otherwise, the seminar report produced under supervision represents a group achievement and forms part of the overall research of the supervising institution. As a result, none of the co-authors (e.g. authors of text, creative project staff, co-supervisors) may use passages from the thesis for commercial purposes or make them accessible to third parties without the (written) approval of all those involved due to reasons of copyright. Particular note must be taken of the Arbeitnehmererfindergesetz (German Employee Invention Act), according to which pre-publication of patent-related content is prohibited.

I agree to submit my written work as an electronic document. I agree that this document will be reviewed using anti-plagiarism software.

I agree to a distribution of my submitted report to all participants of the seminar.

Kiel, February 02, 2024

_____

Abidur Rahman

Abstract

Speech recognition has been researched for decades to enable machines to understand human speech. Initially, only audio speech was considered for speech recognition tasks. Soon, the visual modality was added to the field due to the fact that audio speech recognition systems performed poorly in the presence of additional noise to the acoustic source. By combining both audio and visual features, the performance of the systems improved compared to audio features alone. This paper gives an overview of different approaches to bimodal speech recognition. This paper articulates both audio and video feature extraction algorithm with relative comparison between methods where available. In addition, a brief overview is given of bimodal approaches that could be applied to speech recognition tasks for people with dysarthria disorders. People with dysarthria are unable to control speech articulators properly due to a neurological disorder, resulting in inaccurate pronunciation, slow speech and a low voice. This ultimately affects the quality of speech. In this case, the visual features could come in handy, as they aren't affected by the phoneme labelling imperfections.

1 Introduction

Automatic Speech Recognition (ASR) is a revolutionary system that reliably converts spoken language into written text. ASR systems analyse audio signals using sophisticated algorithms to provide services such as transcription and voice commands. However, in the presence of ambient noise and poor acoustic channels, the performance of an ASR system drops drastically [17], as most ASR systems are used in a laboratory-controlled environment where low noise is ensured.

To overcome these limitations, the visual modality is used, i.e. the video frames are processed in parallel to extract features to increase the speech recognition rate. This is done either by merging the two feature vectors prior to the recognition task, or by combining individual recognition with the associated weight from each modality. The reason for including visual features is that the movement of the mouth muscles provides information about the phonemes [22]. Bimodal *audio-visual speech recognition* (AVSR) is a more robust recogniser because the visual cues are not affected by cross-talk or noise [6], although it has its own limitations.

The aim of this paper is to provide a brief description of different approaches to extract features from audio-video sources, how these features are combined, the models used and the evaluation of each approach. Recently, the use of visual features to improve speech recognition of people with dysarthria, a speech disorder that affects the ability to control speech articulators during speech, has been investigated. A chapter of this paper is devoted to discussing the scope of the visual modality for speech recognition in dysarthric individuals.

The structure of this paper is as follows. Chapter 2 summarises the audio feature extraction methods commonly used in AVSR systems, as mentioned in later chapters. In Chapter 3, algorithms for visual feature extraction are elaborated. The scope for adding visual features to dysarthric speech recognition and the success rate are discussed in Chapter 4. Finally, an outlook of the paper is given in the last chapter.

2 Audio Feature Extraction

2.1 Mel-frequency Cepstral Coefficient

MFCCs are coefficients that represent the short-term power spectrum of a sound signal. These are widely used in speech and audio processing because they effectively capture the spectral characteristics of the signal in a way that is more perceptually relevant to the human ear. To obtain MFCC features from digitised audio input, several successive block operations are required, i.e. pre-emphasis, framing and windowing, calculation of the power spectrum, application of a Mel filter bank, calculation of the logarithmic value of all filter banks, and finally application of the DCT [19, 16] as shown in Fig. 2.1.

Pre-emphasis increases the magnitude of higher frequencies relative to lower frequencies in order to alter the energy distribution across frequencies [16]. This process involves the application of a high pass filter with a pre-emphasis coefficient $\alpha \in [0.95, 0.99]$ according to

$$x^{'}(n) \;=\; x(n) \;+\; \alpha \cdot x(n-1) \tag{2.1}$$

The short-term speech signal is assumed to be stationary and to have stable acoustic characteristics. To achieve this, the whole signal is segmented into several frames of duration $20\text{ms} - 30\text{ms}$. Often the frames are overlapped by 10ms to ensure the temporal characteristic [16]. On each frame, usually either a *Hamming* or*Hanning* window is applied to narrow the signal towards the frame boundary, which is defined by

$$w(n) \;=\; (1-\alpha) \;=\; \alpha \cos\left(\frac{2\pi n}{L-1}\right) \tag{2.2}$$

where, L is the window size.

In the next step, *Discrete Fourier Transform* (DFT) is applied to each windowed frame to extract the spectral information according to the following equation

$$X(k) \;=\; \sum_{n=0}^{N-1} x(n) \exp(-j\frac{2\pi}{N} kn) \tag{2.3}$$

An efficient algorithm called *Fast Fourier Transform* (FFT) is used to compute the DFT by selecting N is a power of 2 and $N \geq L$. This operation results in a *spectogram* where the frequency bins are equally spaced and highly correlated [1] in each frame. To convert this linear frequency scale to the level of human perception, a Mel scale is generated using a set of triangular filters, evenly spaced in the Mel scale. Each filter is applied to the entire spectrum and the result is a set of filterbank coefficients. Typically, 40 filters make up the Mel filterbank [16], and the conversion from linear scale to Mel scale is done according to

$$m \;=\; 2595 \log_{10}\left(1 + \frac{f}{700}\right). \tag{2.4}$$

And the calculation of the power spectrum in the Mel scale for each filter is given by

$$y(m) \;=\; \sum_{k=1}^{N} w_m(k)|x(k)|^2 \tag{2.5}$$

where k is the DFT bin number and m is the Mel filterbank number. Once this is done, *logarithm* is applied to each filterbank energy to make the features less sensitive to acoustic coupling variations and to remove phase information that is not important for speech recognition [1]. Finally, the *Discrete Cosine Transform* (DCT) is applied to the Mel filterbank to select the most accelerating coefficients or to separate the relationship in the logarithmic spectral magnitudes of the filterbank [16]. The DCT expresses a finite sequence of data points in terms of a summation of cosine functions oscillating at different frequencies.

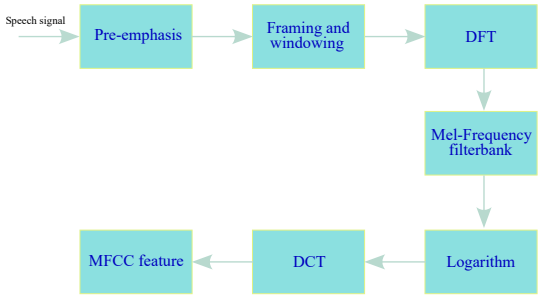

Figure 2.1: Methods of MFCC feature extraction, according to [6].

In [6], a feature vector of length 39 is used to represent the audio frames. The first 13 of the 39 coefficients for each frame are obtained directly from the filter bank; the remaining coefficients are produced by the first and second derivatives. In [14], only the spectrogram is used to generate a 321-dimensional spectral image for each 40ms audio frame. 12 and 14 features from each frame were counted in [11] and [20] respectively.

2.2 ResNet-based Feature Extraction

An audio feature extraction method using *Residual Networks* (ResNets) is presented in [12]. ResNets are a special type of *Convolutional Neural Networks* (CNNs), which establish shortcut connections between layers to solve complex problems that require tons of layers between the input and output layers [2]. It has been experimented that a comparatively deeper 56 layered neural network is more erroneous than a network with 20 layers in both training and testing due to the vanishing gradient problem, meaning that the gradient becomes shallower as it passes through multiple layers during backpropagation [8]. To overcome this, ResNets create a direct path from an input layer to a depth layer and modify the output of the depth layer, $F(x)$, according to

$$H(x) = F(x) + x \tag{2.6}$$

Where x is the input and $H(x)$ is the modified output.

In [12], the standard 18-layer ResNet is used to extract features from the raw audio input by convolution with a 1D kernel. The main reason for using a 1D kernel is that the input is essentially a 1D array. The kernel is of length 5ms with a step of 0.25ms in the first convolution layer, while it is a 3 by 1 array for the remaining convolution layers. The output of the ResNet-18 is fed into a 2 layer *Bidirectional Gated Recurrent Unit* (BGRU) consisting of 1024 cells in each layer as shown in Fig. 2.2. The use case of the BGRU is to further capture the temporal dynamics of the features, i.e. the first and second derivatives. BGRU is a type of recurrent neural network architecture that processes sequential data in both directions to capture information from the past

and future of each time step, providing a more comprehensive understanding of the sequential data.

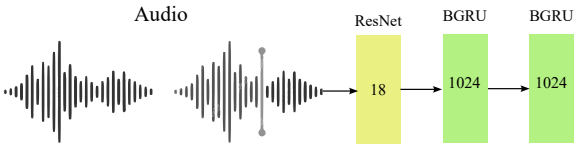

Figure 2.2: Feature extraction from raw audio with ResNet.

2.3 Perceptual Linear Prediction Features

The authors in [13] use an acoustic front-end based on *Perceptual Linear Prediction* (PLP). PLP takes a different approach by directly modelling the human ear's perception of sound through a psychoacoustic model [18]. In PLP-based audio feature extraction, cube-root compression is performed on the Mel filter bank instead of logarithmic compression as in MFCC, as shown in Fig. 2.3. According to [1], the cube root compression is expressed as

$$y(m) = |y(m)|^{1/3} \tag{2.7}$$

This cube root compression represents three perceptual aspects: the critical band resolution curves, the equal loudness curve and the intensity-loudness power law relationship. This non-linear transformation is intended to approximate several perceptual aspects of human hearing. It is used to model the equal-loudness curve and the intensity-loudness power-law relationship, both of which involve non-linearities in human auditory perception [18]. PLP discards irrelevant information from the audio, making the ASR systems slightly better in terms of accuracy and more robust to noise [1, 18].

In [13] the audio stream was first downsampled to 8 kHz. PLP parameters were then computed for every 10 ms on 30 ms sample frames. The full feature vectors consisted of 25 parameters, of which 12 were PLP coefficients. The second set of 12 parameters are the first time derivatives of these coefficients and the last parameter is the energy. Derivatives capture the rate of change of feature values with respect to time. This is particularly relevant for speech recognition, where phonetic transitions and temporal patterns play an important role.

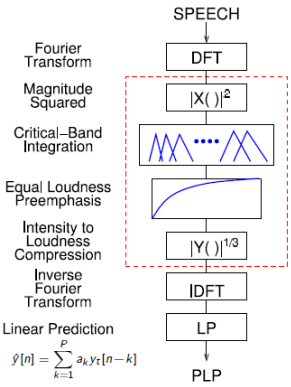

Figure 2.3: PLP-based audio feature extraction approach. [Source:[1]].

2.4 DNN-based Features

A *Deep Neural Network* (DNN) based denoising autoencoder is used to efficiently represent MFCC features from the raw audio input in [9]. Once the MFCC coefficients are computed from the raw audio signal, *Principle Component Analysis* (PCA) is applied to the coefficients. PCA is a non-parametric method for extracting relevant information from redundant features. With minimal additional effort, PCA provides a roadmap for reducing a complex data set to a lower dimension to reveal the sometimes hidden, simplified structure that often underlies it [25]. A deep autoencoder is then trained on the reduced feature vector to learn the bottleneck feature, i.e. the most compressed feature of the input data.

In [9], Gaussian noise is first added to degrade the quality of the audio features. The DNN is then trained in a supervised manner to extract clean features from the artificially degraded audio features. These features are then fed to a GMM-HMM based model [23] for the recognition task. HMM is an abbreviation for *Hidden Markov Model*, which represents the phoneme transition, while GMM stands for *Gaussian Mixture Model*, which models the distribution of the observed feature vector from each state. The method presented in [9] is robust to the additional audio noise and produces clean features from degraded ones. The main purpose of using a GMM-HMM instead of a DNN-HMM is to use the multi-stream HMM as a multimodal integration mechanism for the next AVSR task in an elegant way.

3 Visual Feature Extraction

By using only the audio modality, a speech recognition rate of up to 95% can be achieved in quiet mode [17], but it becomes difficult to recognise the utterance in a noisy environment. This leads to the development of methods to extract features from the angle of the mouth position. So far, several techniques have been developed to extract useful information from the face region that facilitates word or phoneme recognition. Some of these approaches are described below.

3.1 Cepstral Image-based Feature Extraction

A feature extraction method based on *cepstral images* is presented in [7]. First, a sequence of 2D images is considered, representing the evolution of an initial image from the beginning of the sequence, as shown in Fig. 3.1. As the images are 2D, each pixel has its own temporal change. Next, *Linear Predictive Coefficients* (LPC) are generated for each pixel over the time span, where linear prediction assumes that the current pixel intensity can be accurately estimated as a linear combination of past pixel intensities [3] according to

$$\hat{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \tag{3.1}$$

where p is the number of past samples and α_k are the LPC coefficients computed by minimising the sum of the squared differences between the actual speech samples and the linearly predicted ones [3] according to

$$E_n = \sum_{m} e^2(m) \tag{3.2}$$

Where $e(m)$ is the difference between the actual intensity and the predicted intensity.

$$e(n) = s(n) - \hat{s}(n) \tag{3.3}$$

Once the LPC coefficients are obtained, they are converted into cepstral coefficients for each pixel position. By arranging the cepstral coefficients of pixels according to their relative position in an image, a cepstral image is generated for the entire image for any given time span. Then, the 1st and 2nd order cepstral images are obtained from the original one. To extract spatial statistical information from the cepstral images, *Higher Order Local Autocorrelation* (HLAC) of order 2 is applied to each order. Finally, by checking for different spatial correlations of pixel values, 35 coefficients are generated from each order cepstral image.

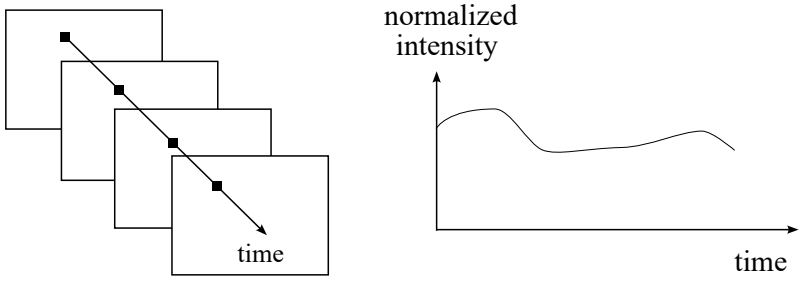

Figure 3.1: Evolution of image pixel intensity [Source:[7]]

3.2 Feature Extraction Using Discrete Cosine Transform

The video processing approaches presented in [6] are quite straightforward and easy to implement. First, the face and mouth region are detected from the whole image using the Viola-Jones algorithm available in the OpenCV library in Python. The Viola-Jones algorithm applies Haar filters to the image to extract Haar-like features for face detection. These filters consist of light and dark regions and produce a single value by subtracting the sum of the pixel values under the light region and the sum of the pixel values under the dark region. The calculation process of this single value requires integral images to reduce the computational complexity. From an original pixel value of an image, the integral pixel value for the integral image is calculated by summing the pixel values above and left to the right (inclusive) of any given pixel position. Once the features are extracted, the *AdaBoost* algorithm is applied to find a set of best features. Finally, a multi-stage cascade classifier is applied to predict the face region [10].

Once the mouth region is detected, the region is resized to make the features invariant to the lip position. The image is then converted from RGB to grey. The DCT is applied to generate a feature matrix of the same size as the input image [6]. For any given discrete signal x_n of length N, the DCT can be calculated according to

$$X(k) = \sum_{n=0}^{N-1} x_n \cdot Cos(\frac{2\pi jnk}{N}) \tag{3.4}$$

Finally, 3 feature vectors of different types could be obtained from the feature matrix by performing three formulas.

First, a number of maximum values (e.g. 3) are taken from the matrix along with their first and second derivatives. This formula gives a feature vector of total 9 values.

The second method applies a 2D window of any shape smaller than the actual feature matrix to the upper left corner and extracts the matrix elements under the window. These elements represent the feature vector.

The third method is a combination of the first and second. First, a window is applied to the upper left corner of the feature matrix, and then a set of maximum values from all the elements under the window are considered for the feature vector representation.

The complete flowchart of the feature extraction according to [6] is shown below.

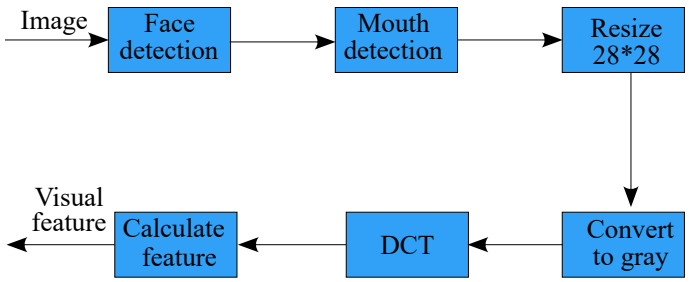

Figure 3.2: Visual feature extraction utilizing DCT according to [6]

.

3.3 ResNet-512 Model-based Feature Extraction

The authors in [14] used almost the same network architecture for visual feature extraction as described in 2.2, except for some minor changes and the use of two different models. One recognition model is *Connectionist Temporal Classification* (CTC) loss based, which predicts the output frame by frame. The CTC loss based model uses only the encoders and includes self-attention layers within the transformer architecture. This requires an external speech model to orient the output words. Language models play a crucial role in correctly orienting words in a sentence by capturing the contextual relationships and dependencies between words. They take only the previous word and an input and predict the next character by performing a left-to-right beam search, allowing more accurate and coherent text generation or understanding [14]. The other is the sequence-to-sequence (seq-to-seq) model, which uses both the encoder and decoder layers to produce output. Here the encoder part processes the input sequence and the decoder part generates a sequence of translated output that incorporates the processed sequence from the encoder layer [4]. As the seq2seq model translates text sequentially, no further language model is required.

To extract the visual features, a CNN-based face detector is used to extract the face region, followed by a colour histogram to track the mouth region. A spatio-temporal visual front-end is then applied to the 112×112 input image. This consists of a 3D convolution layer with a filter width of 5 frames, followed by a 2D ResNets to reduce dimensionality. This visual front-end results in a 512 one-dimensional feature vector for each video frame [14].

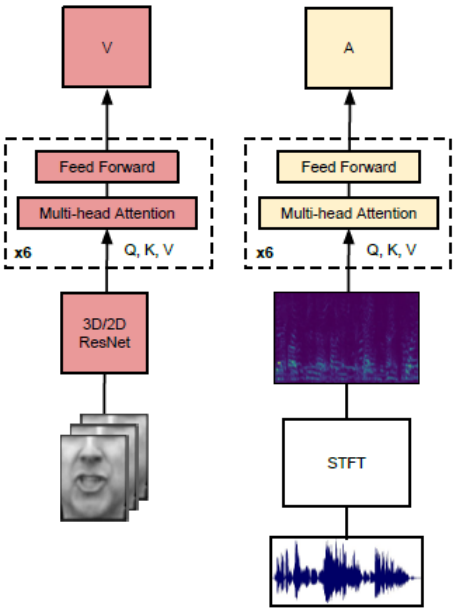

Figure 3.3: 2D ResNet based video feature extraction [Source:[14]].

3.4 Sieve Algorithm Based Features

The manual extraction of the mouth region from each image is shown in [21]. A window of 80×60 pixels was placed in the centre of each image sequence, representing the mouth image.

The subject's mouth never leaves this area, even if the head moves slightly. In other experiments, this step was automated by locating the face using an area sieve algorithm.

A one-dimensional recursive median sieve [21] is applied vertically to the image. The sieve algorithm uses morphological filters to simplify the signal across multiple scales. In this case, it generates a granule of values representing the scale, intensity and position of each pixel value. Granules are essentially sets of pixels that represent features in the image. The intensity value is discarded as the authors intended to consider features independent of intensity variation. Position information is also avoided to maintain a data-driven paradigm, as its use would require identifying and tracking an important set of granules. The scale parameter, which represents granule size, is relatively insensitive to intensity variation and translation. The number of granules of a given scale in the image may change as the scene varies from frame to frame.

For each vertical dimension of 60 pixels of each image, a scale histogram is generated by applying a one-dimensional sieve. A two-stage dimensionality reduction is then performed, where the first stage is a 2-sample moving average that results in 30 coefficients, while the next stage is a PCA analysis that further reduces the feature vectors to 10 values. The resulting feature vector forms the observation vector for a 10 state HMM model for word recognition.

4 Scope for Dysarthric Speech Recognition

Until recently, very little research has been done on dysarthric speech recognition. People suffering from dysarthria are unable to control and coordinate the muscles of the mouth due to neurological motor disorders [5]. This results in low accuracy, low voice and intelligibility of pronounced speech and thus hampers the speaker's ability to communicate with others and devices. Sometimes dysarthria accompanies people with neurological disorders who are physically weakened [24], making them unable to interface with digital devices. In this case, ASR technologies are useful to assist such individuals to communicate with machines and devices, but the phonemes pronounced by dysarthric individuals are quite different from normal speech, which makes it harder for ASR systems to recognise the speech perfectly [24]. In order to recognise the distorted speech, visual features around the mouth articulators come in handy along with audio-only systems. Some notable works on speech recognition for people with dysarthria are mentioned below.

The authors in [5] proposed a two-stage AV-HuBERT framework for dysarthric speech recognition. In this paper, two time frames were presented: pre-training and fine-tuning. In the pre-training frame, features from both audiovisual modalities are extracted and fused and then encoded using a transformer, while in the fine-tuning frame, the audiovisual features are decoded using another transformer. Both time frames are operated in two stages. In the first stage of pre-training, the motor-visual information of the functional areas of the face is fused. The functional areas of the face are divided into five regions: lip (mouth), lower jaw, left and right cheeks, and nose, using a face detector from the *Dlib toolkit*. A CNN-based architecture is then applied to fuse the five different face regions and extract meaningful features. This video feature is used in the second stage, along with audio features, to encode information from both modalities, which ultimately pre-trains the transformer for recognition in the fine-tuning frame. 26 MFCC features are extracted from each acoustic frame and then four consecutive audio feature vectors are integrated into one vector as input. In the fine-tuning frame, features from both modalities are similarly extracted and given as inputs, and the transformer architecture outputs words. The best word error rate (WER) of 6.05% is achieved for mild dysarthric speech, while 63.98% WER is observed for severe dysarthric cases in this experiment.

A comparatively simpler implementation approach is presented in [6], where MFCC features are used for the audio stream as in sec. 2.1 and DCT features are extracted for the visual modality using the Viola-Jones algorithm for mouth region detection as described in sec. 3.2. The early integration model, where features from both modalities are fused early in the processing pipeline resulting in a single set of observations, is used to concatenate both feature vectors [20]. A linear interpolation technique is performed to sample the video frame rate of 30Hz to match the audio features which have a frame rate of 44.1KHz. The authors first evaluated a speaker-dependent AVSR system, where they trained and tested the system using data from five individual speakers with different intelligibility ranges from 39% to 93%. The number of words considered for the evaluation is 2325, consisting of 10 digits (zero to nine), 26 alphabet words, 19 computer commands and 100 common words, each set repeated 3 times. Three methods were used to calculate the length of visual features, as described in Sec. 3.2. Compared to the audio-only recognition rate (57.92%), this speaker-dependent AVSR system performs better only for the most intelligible speaker, with an average accuracy rate of 60.99%. For the speaker-independent method, data sets from all speakers are considered for training and testing, using only 10 digit words with 3 repetitions. Ten speakers were chosen with intelligibility ranging from 29% to 95%. The result is shown only for the 2^{nd} method, which shows that the AVSR system achieves an average accuracy of 60.5%, while the audio-only system achieves an average of 58% for all

speakers.

In [24], a visual data augmentation approach is presented where speech is visually represented by a voicegram. A voicegram is a dynamic representation of the frequency spectrum of a signal that eliminates phoneme-related challenges for dysarthric speakers. The voicegram captures correlations between different dysarthric utterances of the same word that are not seen between time domain waveforms [24]. This results in the elimination of redundancy tasks in ASR systems, such as denoising the waveforms. A 2-dimensional *Spatial Convolutional Neural Network* (S-CNN) is used to train the model by looking at the heat map derived from the voicegraph. The architecture is made up of 8 convolutional layers with 32 filters in the first layer, rising to 256 filters in the last layer. The heatmap is resized to 150×150 pixels for input to the S-CNN. To generate additional voicegram from the existing data, visual data augmentation is used where the extracted voicegrams are modified with different operations. This helps to overcome the data scarcity problem for dysarthric speech recognition. The authors also use a method called transfer learning to maximise the effects of learning from healthy speech visuals. In general, transfer learning uses a pre-trained network to classify new sets of input data by changing the properties of the input and output layers, while freezing the parameter matrix for the rest of the network. In addition, a text-to-speech generation system is used to produce synthetic dysarthric speech to incorporate the existing model for better performance. The evaluation shows that better results are achieved by adding synthetic data with word recognition accuracies of 33.66%, 69.44% and 89.54% under very low, medium and high intelligibility respectively.

The authors of [15] review the clinical research literature to determine what influences the ASR performance of dysarthric speakers when using commercial speech-to-text software, including the fatigue factor during speech, speech variability, misuse factor, other personal factors, and so on.

In the above papers, one approach may perform better than another, but it can't be ignored that in each paper the authors considered different datasets for their purpose, and therefore it is difficult to generalise the acceptability of each method. It is possible that a method that performs better on one dataset may show poor accuracy on another dataset. Although [6] notes that the scope of this experiment is very narrow, only 10 digit words are counted for the recognition task during the speaker-independent experiment. Furthermore, the HMM technique is used, which has recently been replaced by DNN-based models. So it would be a good decision to exclude it from consideration. Both [24] and [5] are DNN-based models and perform quite well on their respective datasets.

5 Conclusions

In this paper, different approaches to feature extraction are presented, along with the advantages and disadvantages of audio and visual modalities. In chap. 4 the scope for dysarthric speech recognition is discussed. Three contributions are selected and their relative performance to the audio-only modality is articulated. It is evident from the above discussions that although state-of-the-art ASR systems demonstrate high performance in quiet environments, the performance is limited by several factors, especially for dysarthric speech where the utterances are severely affected by articulator disorders. In such cases, visual features provide additional support to the system in word recognition tasks. Very few of the available methods for video feature extraction are experimental and tested on a specific dataset, which makes it difficult to obtain a global performance analysis. In the future, it might be a wise decision to apply other methods and to include large datasets for training and testing and to evaluate the results.

List of Figures

Bibliography

[1] Lectures on Speech Signal Analysis. Hiroshi Shimodaira et al. On 17,21 January 2019. URL: https://www.inf.ed.ac.uk/teaching/courses/asr/2018-19/asr02-signal-handout.pdf.

[2] Hussain Mujtaba, "Introduction to Resnet or Residual Network". URL: http://ursula.chem.yale.edu/~batista/classes/CHEM584/Resnet.pdf.

[3] Linear Predictive Coefficients lectures, University of California. URL: https://web.ece.ucsb.edu/Faculty/Rabiner/ece259/digital%20speech%20processing%20course/lectures_new/Lecture%2013_winter_2012_6tp.pdf.

[4] Ashish Vaswani et al. "Attention Is All You Need". In: *Neural Information Processing Systems* (2017).

[5] Chongchong Yu et al. "Multi-Stage Audio-Visual Fusion for Dysarthric Speech Recognition With Pre-Trained Models". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2023).

[6] Elham S. et al. "Audio-Visual Speech Recognition for People with Speech Disorders". In: *International Journal of Computer Applications* (2014).

[7] Eun-Jung Holden et al. "Visual Speech Recognition Using Cepstral Images". In: *Proceedings of the IASTED International Conference On Signal Image Processing* (2000).

[8] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *IEEE* (2015).

[9] Kuniaki Noda et al. "Audio-visual speech recognition using deep learning". In: *Springer Science+Business Media* (2014).

[10] P. Viola et al. "Rapid object detection using a boosted cascade of simple features". In: *IEEE* (2001).

[11] Seyed Reza Shahamiri et al. "A Multi-Views Multi-Learners Approach Towards Dysarthric Speech Recognition Using Multi-Nets Artificial Neural Networks". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2014).

[12] Stavros Petridis et al. "End-to-end Audiovisual Speech Recognition". In: *IEEE International Conference on Acoustics, Speech and Signal Processing* (2018).

[13] Stéphane Dupont et al. "Audio-Visual Speech Modeling for Continuous Speech Recognition". In: *IEEE Transactions on Multimedia* (2000).

[14] Triantafyllos Afouras et al. "Deep Audio-visual Speech Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[15] Victoria Young et al. "Difficulties in automatic speech recognition of dysarthric speakers and the implications for speech-based applications used by the elderly". In: *Assistive technology* (2010).

[16] Zrar Kh. Abdul et al. "Mel Frequency Cepstral Coefficient and Its Applications: A Review". In: *IEEE* (2022).

[17] C. et al. Bregler. "A Hybrid Approach to Bimodal Speech Recognition." In: *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers.* (1994).

[18] Namrata Dave. "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition". In: *INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY* (2013).

[19] S. B. et al. Davis. “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”. In: *IEEE Transactions on Acoustics, Speech and Signal Processing, (ASSP).* (1980).

[20] Timothy J. Hazen. “Visual Model Structures and Synchrony Constraints for Audio-Visual Speech Recognition”. In: *IEEE Transactions on Audio, Speech, and Language Processing* (2006).

[21] lain Manhews et al. “Audio-visual Speech Recognition Using Multiscale Nonlinear Image Decomposition”. In: *Proceeding of Fourth International Conference on Spoken Language Processing* (1996).

[22] I. et al Matthews. “Audiovisual speech recognition using multiscale nonlinear image decomposition.” In: *Proceeding of Fourth International Conference on Spoken Language Processing.* (1996).

[23] Lawrence R. Rabiner. “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”. In: *Proceedings of the IEEE* (1989).

[24] Seyed Reza Shahamiri. “Speech Vision: An End-to-End Deep Learning-Based Dysarthric Automatic Speech Recognition System”. In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2021).

[25] Jonathon Shlens. “A Tutorial on Principal Component Analysis”. In: *CMU School of Computer Science* (2005).