Audio-visual speech enhancement and background noise removal

Han Geng and Chen Li

Group Name: Mariana

Abstract. Audio-visual speech enhancement is a coding process that suprsses the unwanted noise to highlight the useful audio signal from complex audio and video input. Recent development on video processing and audio quality enhancement provides better visual and aural experience, with the involve of deep neural network, audio/video enhancement will be improved compared to traditional method. The goal of this project is to implement a deep audio-visual speech enhancement network, which will be used to remove background noise from a given speaker video with its corresponding audio spectrogram. For research convenience, the video frames will be focused on the lip region only, and the corresponding audio clips used for both training and experiment will be a mix of the original noise-free audio signal and some generated noise signal. The program will initialize with data provided by a pre-trained network on a word-level lip-reading task, who contains labeled speaker videos and their corresponding noiseless audio signals. Our speech enhancement network will take in as input these data, where noise will be added into the audio signals, complete the training process, and then give the audio spectrum with noise removed as output. This method is different from the automatic speech recognition (ASR) who can only identify speech in a relatively noiseless environment. Instead, it can be applied to a more noisy environment where the magnitude of noise is much higher than usual (low signal-to-noise ratio), and it will also work in multi-speaker environments (such as News conference with translater's voice)

1 Introduction

Recent development on video processing and audio quality enhancement gives us better visual and aural experience, and the application of machine learning makes it possible for voice recognition which makes our lives even more convenient. But in the process of training and testing, the input voice signal must be recorded under relatively noiseless environment, otherwise the noise signal in the sound spectrum will cause serious deterioration in the performance, for the reason that the noise signal would add unwanted fluctuation in the spectrum and cause the spectrum to have different characteristics from its original shape, hence the machine would not recognize it as what it should be.

The major themes and applications of Automatic Speech Recognition was discussed in this paper [1] in 2010, and it proposes several approaches and challenges faced by researchers in this field. For example, the context of the speech, the environment quality of the speech and the speakers themselves would influence how well the task could perform. The models proposed supports several types of speech recognition, such as isolated words, connected words, continuous speech and spontaneous speech. In order to recognize these forms of words and speech, a quiet environment must be ensured because for isolated words especially, the recognizers must obtain a test sample that does not have noise or audio signal on either side of the sample window. Imagine using Siri in a crowded station or in strong wind, it may not be able to receive exact orders that you give under such noisy environment. Maybe Siri is not a good example, but it does represent such a situation where the speech recognition system may not function as we want it to be, because of the presence of noises. Other related applications that involve outdoor or noisy activities can also be affected by low signal-to-noise ratio, such as gambling, helicopters and battleground intelligent management systems, etc.

Because of the development of speech recognition and other technical advances, people are having higher and higher demand on audio signal quality, which is often contaminated by numerous sources of natural and artificial noises. In order to improve audio quality and increase voice signal clearness, the development of voice signal enhancement and noise removal is essential.

Various graphic enhancement and noise removal techniques make it easier for us to extract useful information by recovering video or audio clips damaged by multiple source of noise, and the clearer audio signal can then be used in speech recognition and other field of applications.

Multiple advanced approaches to voice enhancement and noise removal have been proposed over the past few years, the most direct approach is using only audio to segregate monaural speech as supposed in [2] in 2009, which is a little bit different from what we are going to propose (remove noise from single-speaker speech), but the core concept is rather similar. [2] uses supervised learning techniques and computational auditory scene analysis to segregate speech recorded in one-microphone scenario. It focused mainly on the noises caused by room reverberation, which cause severe degradation to voice signal quality and is rather hard to resolve because the reverberating noise signal are mainly repetition of ground truth signal but in lower volume, hence it has similar patterns in audio spectrum as ground truth. This makes it hard for researchers to isolate just one channel of clear voice signal, and even harder when there are multiple speakers in the room, each making speeches and causing room reverberation simultaneously. Since the research only uses audio signal, the major resolution technique is to use inverse filtering, but as the paper suggested, this approach is sensitive to room condition, and the result varies for different indoor settings and reverberation conditions. They proposed a method using supervised learning approach to set up several groups for a period of time-frequency (T-F) unit, each contains a set of harmonic characteristics, then they estimate a binary mask for the time-frequency unit, and this mask filters through all the local T-F units and removes those reverberant mixtures whose target energy is weaker than the interference energy. Value 1 in the mask represents that the target is stronger than the interference, and value 0 represents the opposite.

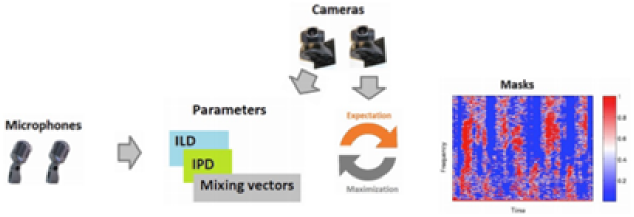


Fig. 1: The direct path parameter vector is calculated with the help of video cameras. (The final probabilistic mask formed from the resulting probabilistic model is used for source separation)

As human beings when we are listening to multiple simultaneous speeches or single-speaker speech with strong noise, we are still able to distinguish the target speaker that we want to hear from other unwanted interfering noises. One reason being that we are using two ears for audio input which gives more characterizing information about the target speaker, but the most helpful and obvious devices that we use to distinguish target speaker is through our eyes.

Since we can observe visual information about our surroundings, we know when the speaker is talking when we see their mouths moving, and from this we know precisely which part of audio clips contains the speech that we desire, and everything happening outside the range of mouth movement is considered noises automatically. With the help of the video information, we can even generate audio signals from a silent video clip if we have trained a model to identify mouth movement and match it to words and sentences as discussed in [3], but this is beyond the scope of the topic of interest. Nevertheless, we can reduce the workload by a great amount and focus on removing noises from only part of the whole audio samples that contains the target speaker's speech, if we have the visual information about the speaker and align it chronologically with the audio samples. The use of corresponding video signals as reference for speaker separation is proposed in [4]. This paper focused on separating single-channeled speakers with help of corresponding visual information.

They record simultaneous speeches given by multiple speakers through a single-channel microphone, and capture videos of these speakers when they are giving the speeches. Using similar techniques as described above, they create a T-F binary mask generated from the visual information that defines which part of the audio is corresponding to the speaker video clip, then filter through the audio signals to retain the dominant speech signals. In this binary mask, value of 1 means the target speaker is talking, and 0 means the speaker is not talking and thus T-F unit is masked. The point of this method is to estimate which part of the T-F component and frequency unit should be retained.

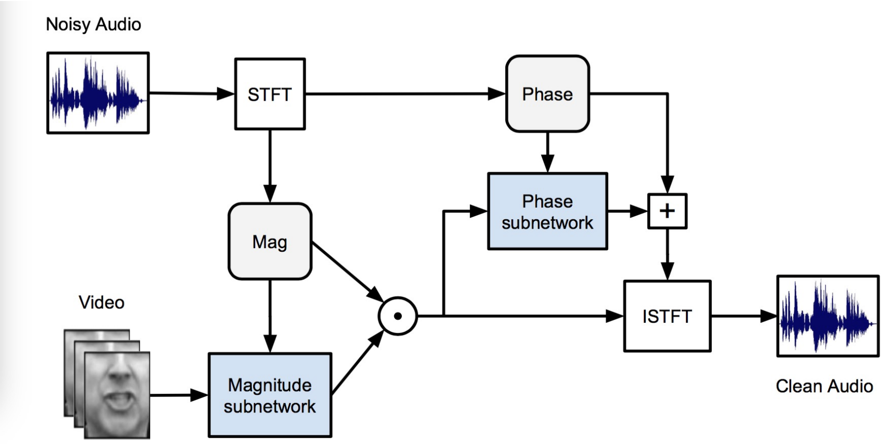


Fig. 2: Audio-visual speech enhancement architecture

This method requires training of a generalized method of moments (GMM) to model audio and visual speech features, so that the correspondence between the audio and visual features can be used to estimate the audio features from the given videos, since this correspondence has been proposed to be valid by several studies in [5] and [6].

In 2017, A. Gabbay et, al. [8] Implemented a model that can extract a specific visible speaker's voice from the similar background noise environments using video data and audio spectrograms. In their study, for unconstrained environments, the separation of the specific speaker requires (i) a sequence of video frames showing the mouth of the speaker; and (ii) a spectrogram of the noisy audio. The source of the audio and video then went through encoder respectively then embedded into fc-layers. Encoding module is composed of a dual tower Convolutional Neural Network in order to make the input from audio/video source have the same embedded representing feature. The video encoder and the audio encoder functions different roles. The video encoder crops from the center of the mouth region of the frames then feed the consecutive gray-scaled cropped frames to consecutive convolution layers (note the layers using Batch Normalization [9] and Leaky-ReLU [10] mathematic matrix processing to filter the frames). The audio encoder receives the input audio spectrogram clips correspond to the video frames fed into the video encoder. The audio encoder also uses convolution layers with Batch Normalization [9] and Leaky-ReLU [10] to filter the audio clips. Both encoders result in the feature vector with the same embedding as output. The two encoders output shared representation to consecutive fully-connected layers. An audio decoder is accepting the data passed through consecutive fully-connected layers with transposed convolution layers to represent the enhanced speech.

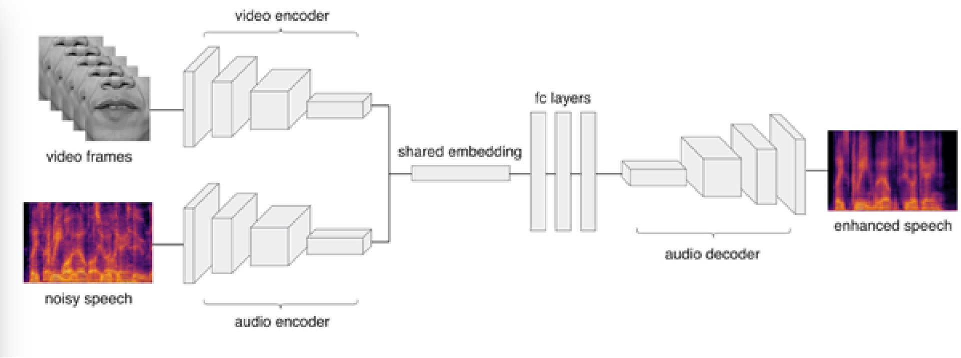


Fig. 3: Audio/video encoder and decoder model for the study below

Aside from the model with video and magnitude of the audio signal example above, [11] in 2017, J.-C. Hou et, al. Introduced a voice enhancement model based on neural network training from a variety of noise features that is able to enhance specific speaker's voice pattern from different kinds of noise backgrounds. The model incorporates audio and visual streams into a unified network model [12]. Using AVDCNN as the audio-visual encoder-decoder network to perform the voice enhancing job. Where AVDCNN is for Audio-Visual Deep Convolutional Neural Network. Similar to the model introduced by A. Gabbay et al., the model introduced by J.-C. Hou et, al. processes audio and visual streams using two independent convolutional neural networks. A fusion network is following the two convolutional neural networks to fuse them together with maximum pooling layer and fully-connected layer [11]. This model uses back-propagation to jointly learn the parameters in an end-to-end manner [12].

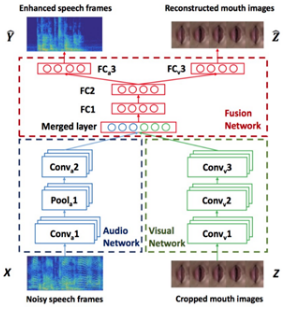


Fig. 4: Construction of the AVDCNN model for the study above

2 Data Set

In order to generate a model for speech recognition and noise removal, we need a large dataset for training and testing the model, one suitable option is VoxCeleb2, which is also used in [7]. It is an open-source media that contains huge amount of audio-visual speaker recognition dataset that we can use in our convolutional neural network training process, which will be discussed in later section. The study proposed in [7] presents a deep CNN based neural speaker embedding system called VGGVox, which is trained to project the voice signal spectrum to Euclidean space, where the Euclidean distance represent the similarity of speakers.

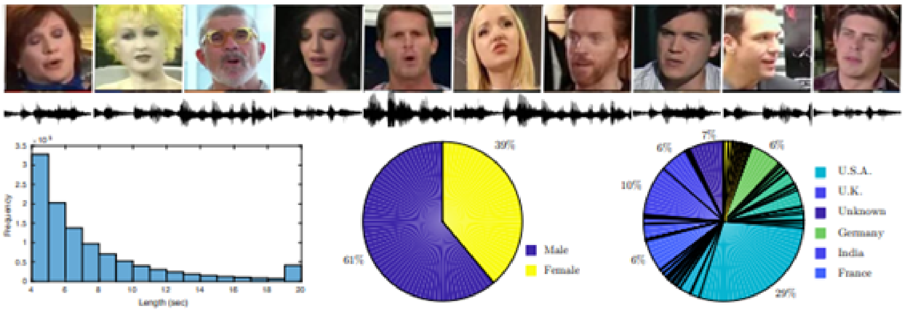


Fig. 5: Top row: Examples of faces from the VoxCeleb2 dataset.
Bottom row: (left) distribution of utterance lengths in the dataset – length shorter than 20s are binned in 1s intervals and all utterances of 20s+ are binned together;
(middle) gender distribution and (right) nationality distribution of speakers. [7]

3 Proposed Method

The algorithm requires audio signal with diversity aspect of noise interacted with the original speaker voice.
In order to create original speaker voice with different kinds of noise mixure we used pysndfx package [21] to add different kinds of noise to the given '.wav' file. The pysndfx package applies audio effects such as reverb and EQ directly to audio files or NumPy ndarrays. The pysndfx package is a lightweight Python wrapper for SoX - Sound eXchange. Where Sox is a cross-platform command line utility that can convert various formats of computer audio files in to other formats. It can also apply various effects to these sound files, and SoX can play and record audio files on most platforms. [22] The pysndfx supports effects range from EQ and compression to phasers, reverb and pitch shifters. Before importing the pysndfx package the terminal should be installed the package related application. (using pip command to install the package related application)

We added mixured level of highshelf, lowshelf filters and reverb, phaser, delay effects to the oriaginal sound and outputed the prcessed sound file and comparsion spectrogram.

A low shelf will either cut or boost signals below a c.o.f. in a manner resembling a shelf, or an equal strength amplitude band past the rolloff. A high shelf filter will either cut or boost signals above a c.o.f. similarly. Where c.o.f stands for cutoff frequency at the point a specific frequency component would have lost approximately half the power (-3 dB) of unaffected frequencies, often referred to as the half-power point. A common synthesis technique is to sweep the cutoff frequency up or down to provide a 'spectral shape' to a sound over time. Cutoff frequencies are usually controlled by an envelope generator or an oscillator (timbre modulation).[23]

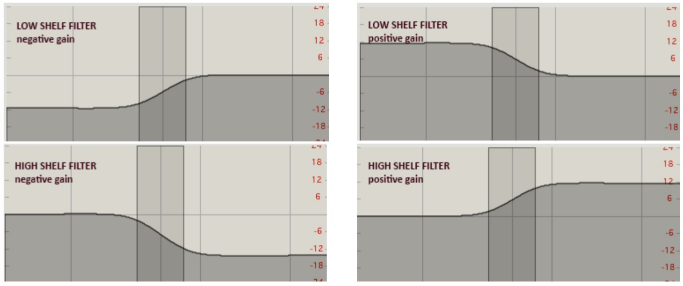


Fig. 6: Shelf transfer function [23]

In detail, The analog transfer function for a low shelf is:

$$H(s) = 1 + \frac{B_0\omega_1}{s+\omega_1} = \frac{s+\omega_1(B_0+1)}{s+\omega_1} \triangleq \frac{s+\omega_z}{s+\omega_1}$$

Fig. 7: Low shelf transfer function [24]

where B_0 is the dc boost amount (at s = 0), and the high-frequency gain (s =) is constrained to be 1 . The transition frequency dividing low and high frequency regions is w1 . See Appendix E for a development of s-plane analysis of analog (continuous-time) filters. [25]
A high shelf is obtained from a low shelf by the conformal mapping (s ¡– 1/s) , which interchanges high and low frequencies.

$$H(s) = 1 + \frac{B_\pi\omega_1}{\frac{1}{s}+\omega_1} = (1+B_\pi)\frac{s+\frac{1}{(1+B_\pi)\omega_1}}{s+\frac{1}{\omega_1}} \triangleq \frac{\omega_z}{\omega_1}\cdot\frac{s+\frac{1}{\omega_z}}{s+\frac{1}{\omega_1}}$$

Fig. 8: High shelf transfer function [24]

In this case, the dc gain is 1 and the high-frequency gain approaches.

$$1 + B_\pi = \omega_z/\omega_1$$

Low and high shelf filters are typically implemented in series, and are typically used to give a little boost or cut at the extreme low or high end (of the spectrum), respectively. To provide a boost or cut near other frequencies, it is necessary to go to (at least) a second-order section, often called a peaking equalizer. [25]

The overview of our proposed method is presented in Fig.9 which has 3 general steps including: 1- Preprocessing; 2- Feature Extraction; and 3- Classification. Next, each step is explained in detail. The goals of CAD system goals are:

1. Extracting the volume based features from the MR T1 images using the automated surface-based analysis package FreeSurfer.
2. Comparing the capability of different types of classifiers for diagnosing PD.

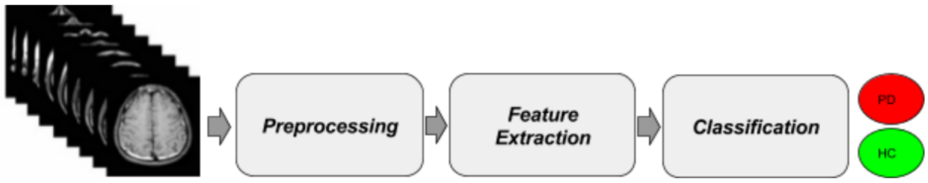


Fig. 9: The general framework of the proposed methods.

3.1 Preprocessing

Preprocessing is an essential step in designing the CAD system which provides informative data for the next steps. In this paper, for computing the volumetric information of the MRI subject's, several preprocessing steps are needed. The Freesurfer image analysis suite is used for performing preprocessing over the 3D MRI data. FreeSurfer is a software package for the analysis and visualization of structural and functional neuroimaging data from cross-sectional or longitudinal studies [16]. The FreeSurfer pipeline performs cortical reconstruction and subcortical volumetric segmentation including the removal of non-brain tissue (skull, eyeballs and skin), using an automated algorithm with the ability to successfully segment the whole brain without any user intervention [17]. FreeSurfer is the structural MRI analysis software of choice for the Human Connectome Project which is documented and freely available for download on-line (http://surfer.nmr.mgh.harvard.edu/). In total 31 preprocessing steps has completed using FreeSurfer, of which some are shown in Fig.10.

There are two types of failures in the preprocessing step which can be categorized into hard failure and soft failure. Hard failures are related to the subjects for whom preprocessing is not successful and the soft failure are related to the subjects that are preprocessed but there are some problem in the results of preprocessing. Out of 568 subjects MRIs, 388 images were successfully preprocessed. Other images were excluded from the dataset due to poor quality of the original images or unknown CDR labels.

3.2 Feature Extraction

After preprocessing using FreeSurfer, a list of volume based features are extracted from different regions of the brain. These features are captured from the regions segmented by brain parcellation using FreeSurfer. Some of the features collected in the left and right hemispheres of the brain are listed below:

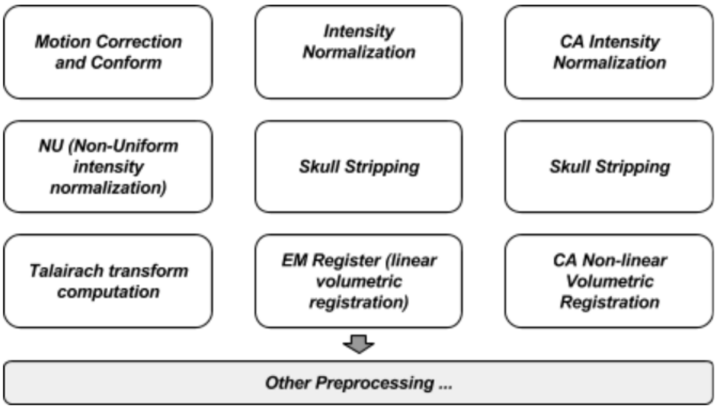

Motion Correction and Conform
Intensity Normalization
CA Intensity Normalization
NU (Non-Uniform intensity normalization)
Skull Stripping
Skull Stripping
Talairach transform computation
EM Register (linear volumetric registration)
CA Non-linear Volumetric Registration
Other Preprocessing ...

Fig. 10: Preprocessing steps.

1. Left and right lateral ventricle
2. Left and right cerebellum white matter
3. Cerebrospinal fluid (CSF)
4. Left and right hippocampus
5. left and right hemisphere cortex
6. Estimated total intra cranial (eTIV)
7. left and right hemisphere surface holes

The extracted feature data is based on Equation 1.

$$FeatureData = \begin{bmatrix} f_{11} & f_{12}\ f_{13} \cdots f_{1n} \\ f_{21} & x_{22}\ x_{23} \cdots f_{2n} \\ \cdots & \\ f_{s1} & f_{s2}\ f_{s3} \cdots f_{sn} \end{bmatrix} \tag{1}$$

Where s is the number of subjects and n is the number of extracted features for that subject. In this study, n is 388 and m is 139.

Furthermore, there are two other type of features which are provided by the PPMI dataset which are age and sex for each subject. Thus, these two biographical information could be added to the extracted feature which gives feature data with the size $(388 * 141)$.

3.3 Classification

In this part the aim is using the extracted volume based features for classifying the MRI data into two classes of PD and HC. In our study, three types of supervised classification algorithm are used. Next, each classification method is described:

- **Logistic Regression (LR):**
  Logistic regression (LR) is a statistical technique which is used in machine

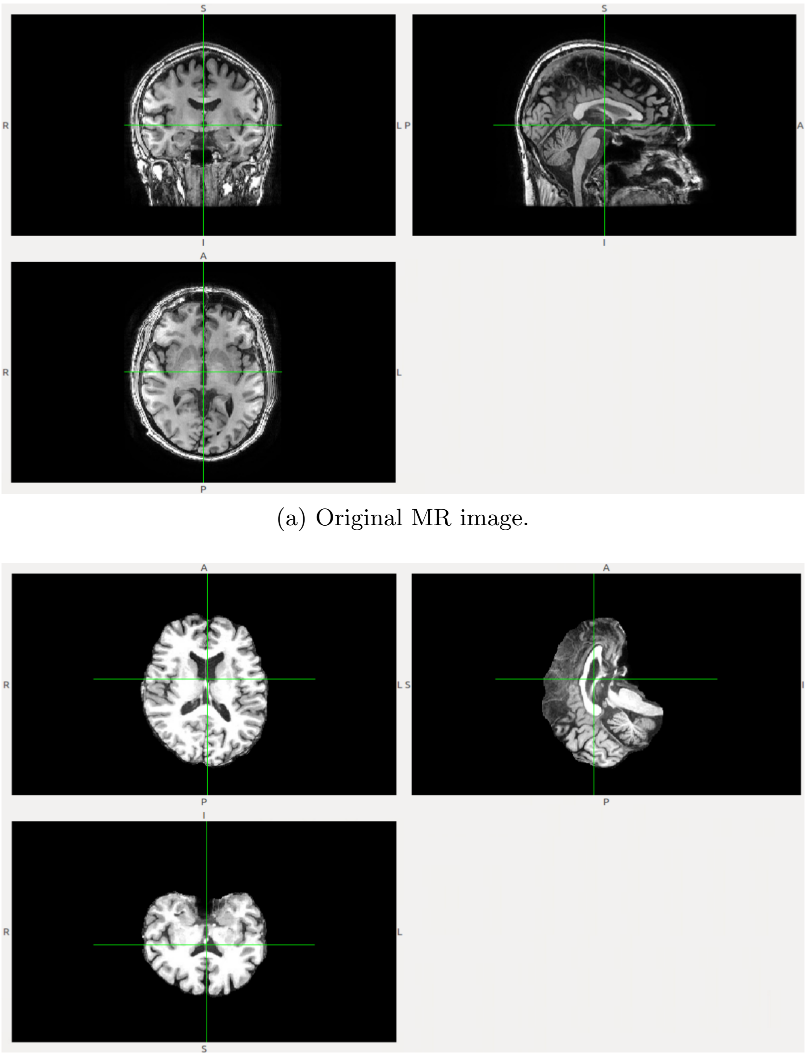
learning for binary classification problems. LR belongs to the family of MaxEnt classifiers known as the exponential or log-linear classifiers [18]. Like naive Bayes, it works by extracting some set of weighted features from the input, taking logs, and combining them linearly (meaning that each feature is multiplied by a weight and then added up) [19]. Thus, this model is a suitable binary classifier for our problem.
- **Random Forest (RF):**
  Random forests (RF) is an ensemble learning method for classification, regression and other tasks. This method is presented by Breiman [20], which creates a set of decision trees from a randomly selected subset of training data. It then aggregates the votes from different decision trees to define final class of the test object. Each tree in a random forest is a weak classifier. A large set of trees trained with randomly chosen data makes a single decision on a majority basis. In the current stage of this research, we tested how accurate decisions can be made by random forests trained by the data coming from a single MRI volume.
- **Support Vector machine (SVM):**
  Support vector machine (SVM) [21] is a well-known supervised machine learning algorithm for classification and regression. It performs classification tasks by constructing optimal hyperplanes in a multidimensional space that separates cases of different class labels. This classification method is more popular because it is easier to use, has higher generalization performance and less tuning comparing to other classifiers. In our case, the kernel SVM is used.

There is a set of parameters for each classifier that needs to be tuned in order to have a fair comparison.

4 Results and Discussion

In this section, the experimental results for different steps of the proposed CAD system for diagnosis of PD is presented. First the preprocessing step prepares the MRI data for the next steps using FreeSurfer. Fig.11 shows the MRI for subject 3102 and the resulting image after preprocessing.

After preprocessing with FreeSurfer, for each subject a list of volume-based features is extracted. Also, age and sex are provided for the PPMI data in their website as demographic information of the patients. Some evaluation has been done over the set of extracted features in terms of their discrimination ability. Since PD is an age related disease, the distribution of data in terms of age feature is plotted. Fig.12 shows the distribution of age in the dataset for the subjects with PD and HC labels. The distribution of all the extracted features are plotted in terms of their ability to distinguish the data into two classes of PD and HC. Some of these distributions are shown in Fig.13. As can be seen in Fig.13(a), the subjects with PD have higher brain volume compared to healthy ones. Furthermore, the distribution in Fig.13(b) and (c) illustrate that when people are in the PD category, their CSF and their CC-Anterior volume size is



(a) Original MR image.

(b) Preprocessed MR image.

Fig. 11: Preprocessing results for one of the subjects.

enlarged. Fig.13(d) shows that the surface hole volume in PD is noticeably higher than the normal subjects. Another set of evaluation is done over the extracted features. Data for every two features are plotted versus each other based on the corresponding class. Fig.14 shows the distribution of data based on the two pair of features including $3rd$ ventricles vs lateral ventricles and $3rd$ ventricles vs. left vessels. In both of them, two features tend to have bigger value when the subject is PD.

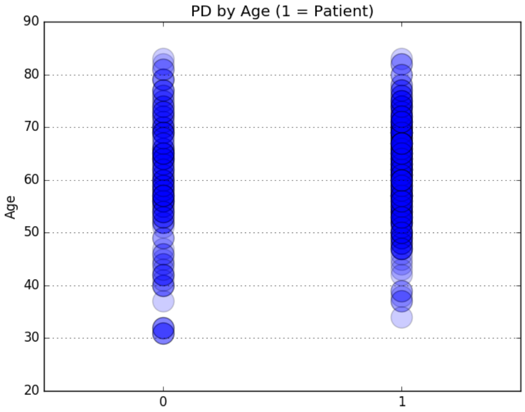


Fig. 12: Distribution of Data in terms of Age feature.

As explained in the previous section, three types of classifiers are used in this study. These algorithm are run over 388 samples with 141 features. Internal and external cross validation is applied with $K = 10$ for external and $k = 5$ for internal (parameter tunning cross validation). The number of selected samples for the training part is 350 and for the test part is 38. Furthermore, the number of PD and HC in each group is presented in Table 1.

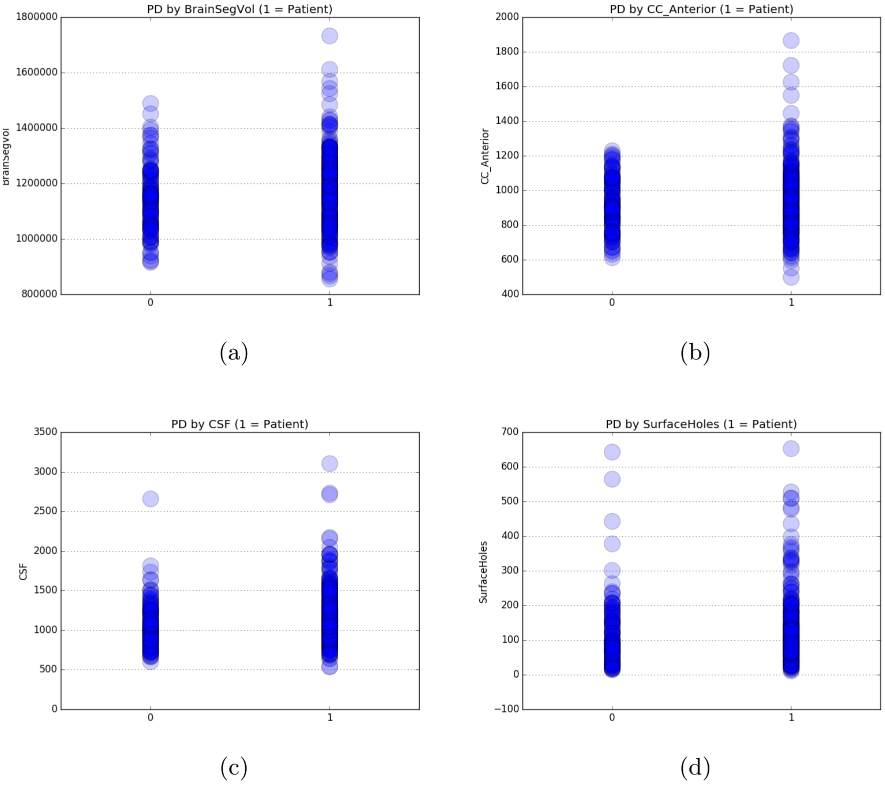
Table 1: Data balance in training and testing parts.

	PD	Hc	Total
Training	236	114	350
Test	26	12	38

As mentioned before, the classification algorithm needs a set of parameters for tunning which is selected as follow:

- logistic Regression (LR):
  Regularization $= [1e-1, 1e-2, 1e-3, 1e-4, 1e-5]$, Tolerance = [1e-1, 1e-2, 1e-3, 1e-4, 1e-5]
- Random Forest (RF):
  Number of estimator $=[5, 10, 15, 20, 25]$, Max depth $= [2-10]$
- Support Vector Machine (SVM):
  C $= [0.1, 1, 10, 100, 1000]$, Gamma $= [10, 1, 1e{-}1, 1e{-}2, 1e{-}3, 1e{-}4]$, kernels $= [linear, rbf, poly]$

The evaluation metrics used in this paper for comparing the results of the classification algorithms include accuracy, confusion matrix (recall, precision) and



(a)

(b)

(c)

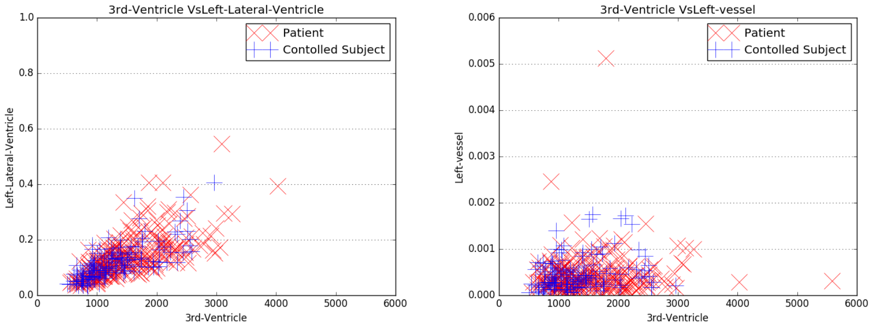
(d)

Fig. 13: Data distributions in terms of the class labels and corresponding features; which are: (a) Brain segmented volume. (b) CC- Anterior. (c) CSF. (d) Surface holes.

AUC (area under ROC curve). The classification results for LR, RF and SVM are shown in Tables 2, 3, and 4, respectively.

Table 5 shows the general comparison between these methods. The best result is for RF. In the table there are two sets of results related to using age/sex feature or doing the classification based only on the extracted volume based features from FreeSurfer.

Based on the literature review, most papers use VBM for data analysis and feature extraction. In this paper, one of the important goals was evaluating the FreeSurfer features for PD MRIs using machine learning techniques. Generally, the experimental results show that the classification models need more information about the data that should be added to the current features, since these are low-level features and we need a set of high-level features as well. In future research, we are going to determine the useful general features that can be combined with the volume based extracted features.



(a)

(b)

Fig. 14: Data distribution based on the pair of features: (a)3rd ventricle and left lateral ventricle. (b) 3rd ventricle and left vessel.

Table 2: Logistic regression performance

Logit	Train Accuracy	Test Accuracy	TN	FP	FN	TP	AUC
K_0	0.6867	0.6500	3	10	4	23	0.5413
K_1	0.6781	0.6750	0	13	0	27	0.5000
K_2	0.6819	0.7179	2	11	0	26	0.5769
K_3	0.6733	0.6666	3	10	3	23	0.5576
K_4	0.6991	0.5384	2	11	7	19	0.4423
K_5	0.6618	0.6153	0	13	2	24	0.4615
K_6	0.6657	0.6842	1	11	1	25	0.5224
K_7	0.7057	0.6578	2	10	3	23	0.5256
K_8	0.6857	0.6052	1	11	4	22	0.4647
K_9	0.6742	0.6315	0	12	2	24	0.4615

5 Conclusion

We presented an automatic MRI based CAD system for diagnosing Parkinson's Disease (PD) which is the second common neurodegenerative disease affecting elderly people. This disease is exposed by loss of neuro-transmitters that control body movements and there is no cure other than earlier diagnosis with better and more efficient treatment for patients. MR T1 images from the public PPMI PD data set is used. FreeSurfer is used for feature extraction and preprocessing. The decision model for classification of the extracted feature data is based on LR, RF and SVM methods. In the experimental results, the ability of these three types of classifiers for PD diagnosis are compared to each other. Results show that using only MRI is a potential option for PD diagnosis. This approach will avoid exposing the brain to harmful radiation based scans. In future work, the efficiency of the proposed method could be improved by adding high level features to the

Table 3: Random forests performance

RandomFrest	Train Accuracy	Test Accuracy	TN	FP	FN	TP	AUC
K_0	0.7270	0.6750	0	13	0	27	0.5000
K_1	0.7327	0.6500	1	12	2	25	0.5014
K_2	0.7507	0.6923	1	12	0	26	0.5384
K_3	0.7392	0.6923	1	12	0	26	0.5384
K_4	0.7335	0.6666	2	11	2	24	0.5384
K_5	0.7277	0.7179	2	11	0	26	0.5769
K_6	0.7399	0.6578	0	12	1	25	0.4807
K_7	0.7171	0.6578	0	12	1	25	0.4807
K_8	0.7257	0.6315	1	11	3	23	0.4839
K_9	0.7199	0.7105	1	11	0	26	0.5416

Table 4: Support Vector Machine performance

SVM	Train Accuracy	Test Accuracy	TN	FP	FN	TP	AUC
K_0	0.7528	0.6000	3	10	6	21	0.5042
K_1	0.7471	0.6000	7	6	10	17	0.5840
K_2	0.7134	0.6923	5	8	4	22	0.6153
K_3	0.7335	0.5128	2	11	8	18	0.4230
K_4	0.7478	0.6666	3	10	3	23	0.5576
K_5	0.7449	0.5897	4	9	7	19	0.5192
K_6	0.7228	0.7105	4	8	3	23	0.6089
K_7	0.7400	0.4473	3	9	12	14	0.3942
K_8	0.7171	0.5789	4	8	8	18	0.5128
K_9	0.7428	0.5789	3	9	7	19	0.4903

current ones. The classification rate with MRI needs to be improved to get close to those using raditation based scanning.

References

1. M. Anusuya and S. K. Katti, “Speech recognition by machine, a review,” arXiv preprint arXiv:1001.2267, 2010.
2. Z.Jinand D.Wang,“A supervised learning approach to monaural segregation of reverberant speech,” IEEE Transactions on Audio, Speech, and Language Processing, 2009.
3. A. Ephrat, T. Halperin, and S. Peleg, “Improved speech reconstruction from silent video,” in ICCV 2017 Workshop on Computer Vision for Audio-Visual Media, 2017.
4. F.KhanandB.Milner,“Speaker separation using visually-derived binary masks,” in AVSP, 2013
5. H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, “Quantitative association of vocal tract and facial behavior,” Speech Communication, vol. 26, no. 1, pp. 23–43, Oct. 1998.
6. I. Almajai and B. Milner, “Visually-derived Wiener filters for speech enhancement,” IEEE Trans. Audio, Speech and Language Processing, vol. 19, no. 6, pp. 1642–1651, Aug. 2011.

Table 5: Comparing performance of different classifiers

Methods/Criteria	Age/Sex Feature	Train Accuracy	Test Accuracy	AUC
LR	No	0.6806	0.6467	0.4912
LR	Yes	0.6858	0.6313	0.4794
RF	No	0.7425	0.67	0.4647
RF	Yes	0.7396	0.6673	0.5086
SVM with rbf kernel	No	0.6752	0.6753	0.500
SVM with rbf kernel	Yes	0.6752	0.6753	0.500
SVM with linear kernel	No	0.7362	0.5977	0.521
SVM with linear kernel	Yes	0.7259	0.5927	0.5273

7. J. S. Chung, A. Nagrani, , and A. Zisserman, “VoxCeleb2: Deep speaker recognition,” arXiv preprint arXiv:1001.2267, 2018.
8. A. Gabbay, A. Shamir, and S. Peleg, “Visual Speech Enhancement using Noise-Invariant Training,” arXiv preprint arXiv:1711.08789, 2017.
9. S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICML’ 15, pages 448–456, 2015.
10. A.L.Maas,A.Y.Hannun,andA.Y.Ng.Rectifier nonlinearities improve neural network acoustic models. In ICML’ 13, volume 30, 2013.
11. J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, “Audio-Visual Speech Enhancement Using Multi- modal Deep Convolutional Neural Networks,” IEEE Transactions on Emerging Topics in Computational Intelligence, 2018.
12. J.-C. Hou, S.-S. Wang, Y.-H. Lai, J.-C. Lin, Y. Tsao, H.-W. Chang, and H.-M. Wang. Audio-visual speech enhancement based on multimodal deep convolutional neural network. arXiv:1703.10893, 2017.
13.
14.
15.
16.
17.
18.
19.
20.
21. PyPI official document-pysndfx 0.3.6: https://pypi.org/project/pysndfx/
22. Github: python-audio-effects. https://github.com/carlthome/python-audio-effects/blob/master/README.md
23. Introduction to Computer Music: University of Indiana http://www.indiana.edu/ emusic/etext/synthesis/chapter4$_{f}ilters.shtml$
24. U. Zölzer, Digital Audio Signal Processing, New York: John Wiley and Sons, Inc., 1999.
25. J. O. Smith, Introduction to digital filters: with audio applications. W3K Publishing, 2008.