


RESEARCH

Open Access

A CNN-based approach to identification of degradations in speech signals



Yuki Saishu¹, Amir Hossein Poorjam^{1,2*}  and Mads Græsbøll Christensen¹

Abstract

The presence of degradations in speech signals, which causes acoustic mismatch between training and operating conditions, deteriorates the performance of many speech-based systems. A variety of enhancement techniques have been developed to compensate the acoustic mismatch in speech-based applications. To apply these signal enhancement techniques, however, it is necessary to know prior information about the presence and the type of degradations in speech signals. In this paper, we propose a new convolutional neural network (CNN)-based approach to automatically identify the major types of degradations commonly encountered in speech-based applications, namely additive noise, nonlinear distortion, and reverberation. In this approach, a set of parallel CNNs, each detecting a certain degradation type, is applied to the log-mel spectrogram of audio signals. Experimental results using two different speech types, namely pathological voice and normal running speech, show the effectiveness of the proposed method in detecting the presence and the type of degradations in speech signals which outperforms the state-of-the-art method. Using the score weighted class activation mapping, we provide a visual analysis of how the network makes decision for identifying different types of degradation in speech signals by highlighting the regions of the log-mel spectrogram which are more influential to the target degradation.

Keywords: Signal enhancement, Convolutional neural network, Identification of degradation, Quality control, Visualization

1 Introduction

Advances in portable devices such as smartphones and tablets, that are equipped with high-quality microphones, facilitate capturing and processing speech signals in a wide range of environments. However, the quality of the recordings is not necessarily as expected, as they might be subject to degradation. In practice, the presence of degradation during the operating time can deteriorate the performance of speech-based systems, such as speech recognition [1], speaker identification [2], and pathological voice analysis (assessment of voice signal of a speaker with a voice disorder) [3, 4], mainly due to acoustic mismatch between training and operating conditions. The

most common types of degradation typically encountered in speech-based applications are background noise, reverberation, and nonlinear distortion.

A speech signal degraded by additive noise, reverberation, and nonlinear distortion can be, respectively, modeled as follows:

$$x_n(t) = s(t) + e(t), \quad (1)$$

$$x_r(t) = s(t) * h(t), \quad (2)$$

$$x_d(t) = \psi(s(t)), \quad (3)$$

where t is the time index, $s(t)$ is the clean speech signal recorded by a microphone in a noise-free and non-reverberant environment, $e(t)$ is an additive noise, ψ represents a nonlinear function, $h(t)$ is a room impulse response (RIR), and the $*$ indicates the convolution operation. We note that in reality, these degradations are even

*Correspondence: ahp@create.aau.dk

¹Audio Analysis Lab, CREATE, Aalborg University, Rendsburggade 14, 9000 Aalborg, Denmark

²Verisk Analytics, 388 Market Street, 94111 San Francisco, CA, USA

more complex. For example, they may be time-dependent. A variety of effective signal enhancement techniques have been developed to enhance a degraded speech signal such as noise reduction [5, 6], dereverberation [7, 8], and restoration of some types of nonlinear distortion [9, 10]. Most of these enhancement algorithms have been designed to deal with a specific type of degradation in a signal, although recent research in comprehensive speech enhancement, dealing with both additive noise and reverberation, is promising [11–13]. Nevertheless, to properly compensate for the effects of degradations, it is necessary to know or obtain information about the presence and the type of degradations in speech signals. Since manual inspection of the signals is very time consuming, costly, and even impossible in many speech-based applications, an accurate degradation detection system would be useful to automatically identify the presence and type of degradations.

There are a variety of approaches to identify different types of degradation in speech signals. For example, Ma et al. in [14] proposed a hidden Markov model-based approach to distinguish different types of noise in speech signals. In another study by Desmond et al. [15], the reverberant signals are detected using a channel-specific statistical model. In [16, 17], clipping in speech signals, as an example of nonlinear distortion, is detected. Although effective, these approaches are focused on detecting a single, specific type of degradation. The use of a multiclass classification, on the other hand, can be used to detect different types of degradations. In [18, 19], Poorjam et al. proposed two generalized multiclass classification-based approaches detecting various types of degradation, which investigated only on pathological voice signals and the accuracy was still inadequate. Moreover, there is no control over the class assignment in these approaches when a new type of degradation is observed for which the classifier has not been trained. For example, clipping, packet-loss, dynamic range compression, automatic gain control, and distortions due to using low quality or improperly configured equipment are considered as new types of degradation for a multiclass classifier trained only with noisy and reverberant signals.

To overcome the limitations of the multiclass-based approaches, one can use a multilabel classification approach in which more than one class labels may be assigned to each sample. Compared to the multiclass-based methods, this approach can better deal with some challenging cases such as the presence of a new degradation type and when more than one degradation coexists. In the former case, the sample may be classified as none of the target classes. In the latter case, more than one detector can accept a signal subject to a mixture of degradations. One possible solution is to integrate the existing algorithms, developed for detecting each

type of degradation, into a unified framework and consider each subsystem as a detector to make a decision about a signal. However, algorithms that are independently developed may make very different assumptions and may have diverse requirements that could occasionally be conflicting. Thus, integrating them into a framework is very challenging, and meeting all requirements at the same time might not be feasible in some cases.

As an alternative solution, Poorjam et al. proposed a data-driven approach which uses a set of parallel Gaussian mixture models (GMMs) to detect three types of degradation in pathological voice signals, namely background noise, reverberation, and nonlinear distortion [4]. All detectors in this approach are similar in terms of the complexity, underlying assumptions, and the acoustic features except that they are trained using different degraded signals. This approach is focused on pathological voices and, particularly, on the sustained vowels.

In this paper, we propose a more accurate convolutional neural network (CNN)-based approach which can identify degradations not only in sustained vowels, but also in normal running speech. CNNs are computationally efficient deep neural networks that are able to learn complex patterns in the spectrogram of a speech signal. In this approach, we apply a set of parallel CNNs to the log-mel spectrograms of the signals. Each CNN model, trained with signals corrupted by a specific degradation type, is responsible for detecting the corresponding degradation in a test signal. The prediction scores of an unseen test sample can be used to associate multiple degradation labels to an observation and can be interpreted as the degree of contribution of each degradation in a degraded signal. Moreover, using the score class activation mapping (score-CAM) technique [20], we visually explain on what basis the CNN models make a specific decision in detecting different types of degradation by finding the regions in the mel-scale spectrograms of a degraded signal that are most influential to the scores of the target class. In this technique, different activation maps are applied to the input spectrogram, each perturbing a region of the spectrogram. Then, the effect of each activation map on the prediction scores is observed. The importance of each activation map is determined by the prediction score on the target class. Finally, a saliency map is generated by a weighted linear combination of all activation maps to visualize the internal representation in a CNN [20]. Since this technique does not require any modifications to the architecture of the network, it can be applied to a wide variety of CNN models.

The rest of this paper is organized as follows. In Section 2, we formulate the problem of automatic degradation detection, and describe the proposed approach. The experimental setup is explained in Section 3. In

Section 4, we present and discuss about the results. The paper ends with conclusions in Section 5.

2 System description

2.1 Problem formulation

In the problem of degradation detection in speech signals, we are given a set of training data $\Lambda = \{\mathbf{x}_n, y_{n,d}\}_{n=1}^N$, where $\mathbf{x}_n \in \mathbb{R}^k$ denotes the n th observation of k dimension. Depending on the system and the level of processing, this could represent acoustic features of an audio signal or a frame of a signal. For example, in the proposed system, introduced in Section 2.2, \mathbf{x}_n represents the log-mel spectrogram of the n th audio signal, and in the baseline system, described in Section 2.3, it is the mel-frequency cepstral coefficients of the n th frame of a signal. $y_{n,d} \in \{0, 1\}$ denotes whether the n th observation belongs to a degradation class d . N is the total number of training samples. The goal is to approximate a binary classifier function g_d for each degradation type d , such that for an observation not in the training data, \mathbf{x}_{test} , the probability of the test sample classified in the correct class is maximized. In other words, the estimated degradation label, $\hat{y}_d = g_d(\mathbf{x}_{\text{test}})$, for $\hat{y}_d \in [0, 1]$, is as close as possible to the true label.

2.2 The proposed method

In our proposed method, we use a set of parallel CNNs to approximate the functions g_d . Each CNN, inspired by VGGNet [21], consists of several convolutional blocks,

and each block consists of several convolutional layers with kernel size of 3×3 . As shown in Fig. 1, we propose 5 different CNN architectures for each detector to investigate the optimal architecture for degradation detection problem. The numbers in front of “Conv” in each layer show the number of feature maps. The CNN32, which has 28,807 parameters to train, consists of one convolutional block of 3 layers. The CNN64, with 120,423 parameters, consists of a 2-layer and a 3-layer convolutional blocks. The CNN128 comprises two 2-layer and one 3-layer convolutional blocks. The number of parameters of this network is 469,543. In CNN256, there are two 2-layer and two 3-layer blocks and has 1,979,175 parameters. Finally, the CNN512, which consists of 7,947,559 parameters, is made of three 2-layer and two 3-layer blocks. In Fig. 2, the architecture of CNN128 is illustrated in more detail. To connect the convolutional layers, we employ batch normalization (BN) and rectified linear unit (ReLU). BN permits a deep neural network to learn with larger learning rates which facilitates quicker convergence and better generalization [22]. The output layer consists of two dense layers—also known as the fully connected layers—that are connected to the last convolutional layer by a global average pooling. We use a sigmoid activation function in the output layer to produce a score in a range $[0, 1]$.

As the acoustic feature, we use the log-mel spectrogram of size 300 frames \times 40 mel bins, calculated by taking the logarithm of the output of a mel-scale filter bank applied to the short-time Fourier transform (STFT) of a signal.

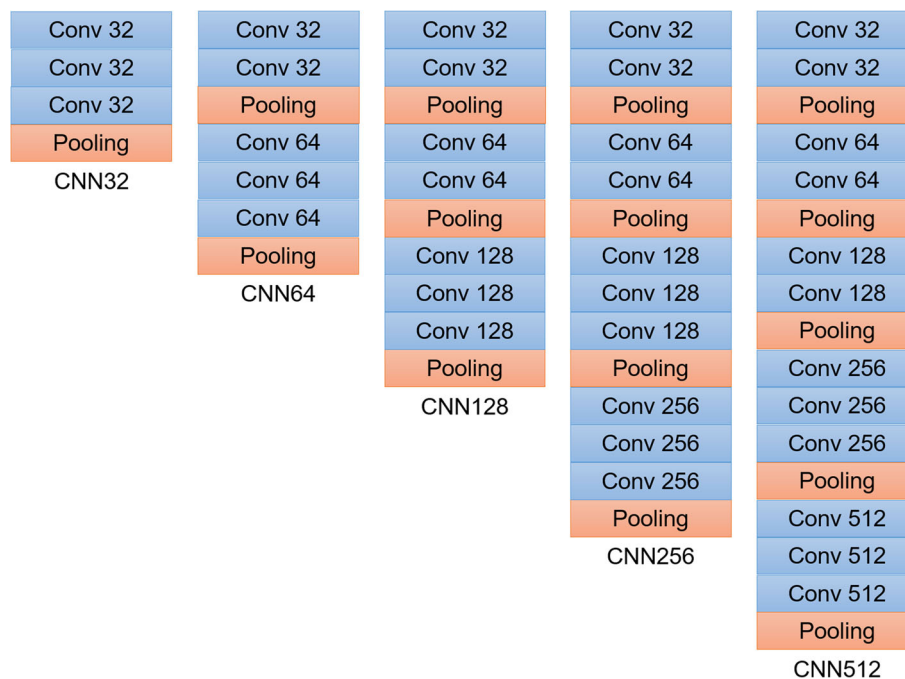
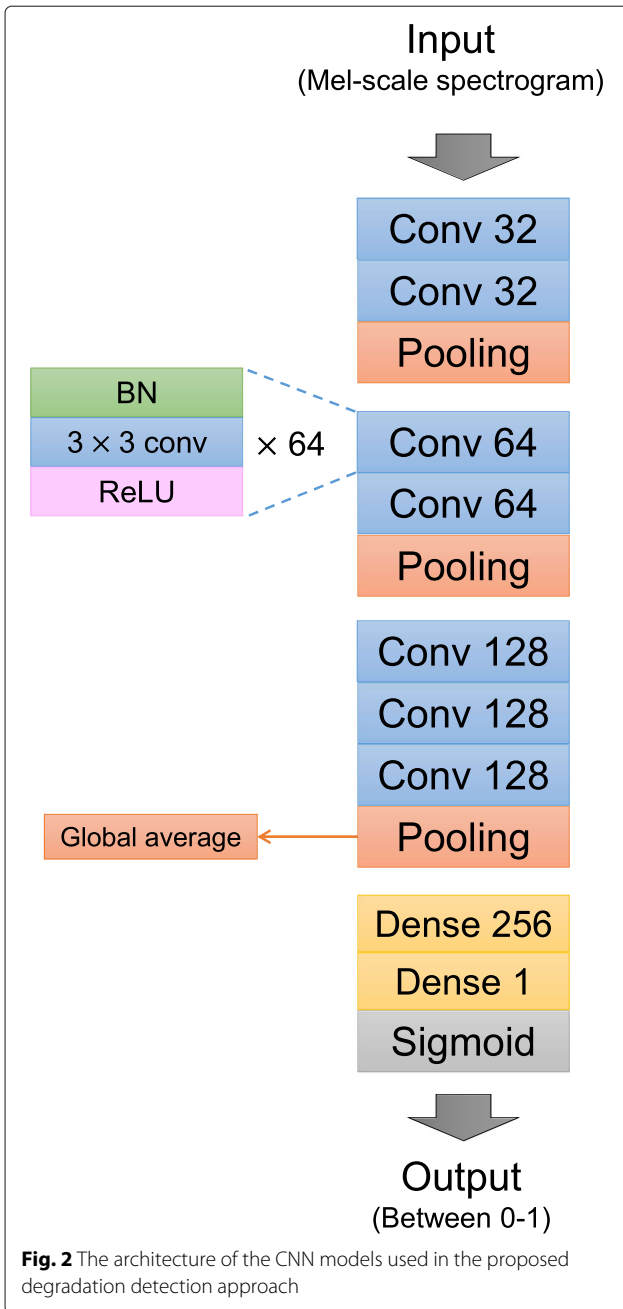


Fig. 1 The architecture of CNN models of different number of convolutional layers



The log-mel spectrogram is a popular signal parametrization technique in many audio applications using deep neural networks which provides an efficient, perceptually relevant, 2-dimensional representation of an audio signal. Compared to the STFT, the log-mel spectrogram provides a less redundant representation of an audio signal and allows CNNs to learn with a smaller number of training data. The decibel scaling is motivated by the human perception of loudness [23] and has shown to provide a better discriminability compared to the linear version [24]. The resulting log-mel spectrogram together with the first- and

second-order derivatives is used as the input feature to the CNN.

We use stochastic gradient descent (SGD) to minimize the binary cross-entropy for each classifier that is defined as:

$$L_d = -\frac{1}{N} \sum_{n=1}^N (y_{n,d} \ln(g_d(\mathbf{x}_n)) + (1 - y_{n,d}) \ln(1 - g_d(\mathbf{x}_n))), \quad (4)$$

where $g_d(\mathbf{x}_n) \in [0, 1]$ is the output score of the CNN trained to identify a specific type of degradation, and $y_{n,d} \in \{0, 1\}$ is the true degradation label.

The decision for the test observation is made by setting a threshold over the output scores of each CNN. This way, if a test sample is subject to a new type of degradation, we expect it to be rejected by all CNNs based on a pre-defined threshold. Moreover, if an observation is subject to more than one type of degradation, we expect that the output score of more than one CNN to be above the threshold. It should be noted that since the selection of an optimal decision threshold depends on the application, in this study, we consider the soft scores and use a threshold-independent metric, introduced in Section 3.4, to evaluate the performance of the proposed system.

2.3 Baseline system

As a baseline system, with which we compare our proposed system, we use the Gaussian mixture model-universal background model (GMM-UBM) degradation detection approach proposed in [4]. In this approach, a set of parallel GMMs, fitted to the frames of the speech signals in the mel-frequency cepstral coefficient (MFCC) domain, is used to detect different types of degradation. The training phase consists of two steps: (1) training a degradation-independent GMM with a very large amount of training data from various degradation classes, referred to as the UBM, and (2) training a set of degradation-dependent GMMs by adapting the parameters of the UBM using the corresponding training data. For evaluation, the identification score of a certain type of degradation, d , and time-sequence input, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n, \dots, \mathbf{x}_N)$, is computed by the following equation:

$$\begin{aligned} \sigma_d &= g_d(\mathbf{X}) \\ &= \frac{1}{N} \left(\sum_{n=1}^N \log p(\mathbf{x}_n | \lambda_d) - \sum_{n=1}^N \log p(\mathbf{x}_n | \lambda_{ubm}) \right), \end{aligned} \quad (5)$$

where the N is the total number of time-frames, the λ_{ubm} and the λ_d are the parameters of the UBM and the degradation-dependent GMMs, respectively, and $p(\mathbf{x}_n | \lambda)$ is the Gaussian probability density function. The identification is made by setting a threshold over the scores.

3 Experimental setup

3.1 Data sets

Our approach can be applied to any type of speech, such as normal running speech, whispered speech, emotional speech, sustained vowel phonation, and singing voice. In this study, we consider two types of speech, namely pathological voice and normal running speech, to evaluate the performance of the proposed method. For the pathological voice, we used the mPower mobile Parkinson's disease (MMPD) data set [25] which includes more than 65,000 voice samples of 10 seconds sustained phonations of the vowel /a/ recorded at 44.1 kHz sampling frequency by PD patients and healthy speakers. This data set has been selected because most PD patients suffer from some form of vocal disorders [26]. Moreover, since sustained vowel phonations provide a simple acoustic structure to characterize the glottal source and resonant structure of the vocal tract [27], they are considered as the main voice material for analysis of pathological voice caused by a range of medical disorders. For the normal running speech, we used an English speech database published by the Center for Speech Technology Research at University of Edinburgh [28]. The samples of this database were recorded at 48 kHz.

3.1.1 Pathological voice

To prepare data for degradation detection experiments in pathological voices, we randomly selected 9,000 samples from the MMPD data set, and divided them into 5 equal groups of 1,800 samples. The recordings of the first group were degraded by six different types of additive noise, namely babble, street, restaurant, office, white Gaussian, and wind noises¹ under different signal-to-noise ratio (SNR) conditions ranging from -10 dB to 20 dB. The noise signals were resampled to 44.1 kHz before being added to the voice signals. To reduce the probability of observing signals degraded by exactly the same noise segments in both training and evaluation subsets, we added a random segment of a noise file to each clean signal.

The recordings of the second group were filtered by 46 real room impulse responses (RIRs) of the AIR database [29], measured with a mock-up phone in hand-held and hands-free positions in various realistic indoor environments, such as a meeting room, a corridor, a lecture room, an office, a stairway, and a kitchen, to produce reverberant samples. The reverberation time of the RIRs, RT_{60} , defined as the time it takes for a switched-off sound to decay by 60 dB [30], ranges from 390 ms to 1.47 s. The

direct to reverberant energy ratio of the RIRs ranges from 4.35 to 12.28 dB. The RIRs were resampled to 44.1 kHz prior to convolution.

The samples of the third group were distorted by either clipping, coding, or clipping followed by coding as an example of nonlinear distortion. The clipping level, defined as a proportion of the peak absolute signal amplitude to which the sample values greater than this threshold are limited, was set to 0.3, 0.5, or 0.7, and we used 9.6 kbps and 16 kbps code-excited linear prediction (CELP) codecs [31].

We used the fourth group for a combination of additive noise and reverberation, in which a voice sample was filtered by a RIR and added to a noisy signal that was also convolved with a RIR. The noisy signals in this case are degraded by indoor environment noises such as babble, restaurant, and office noise under 0 dB, 5 dB, or 10 dB SNR conditions. The reason for choosing this subset is to evaluate whether a signal, in which both noise and reverberation coexist, can be detected by both noise and reverberant detectors. The fifth group was used without any processing and considered as the clean class.

3.1.2 Normal running speech

To prepare samples for noisy and noisy-reverberant classes in normal speech, we used the clean and noisy parallel speech data set (NS) [28] and clean and noisy-reverberant speech data set (NRS) [32], respectively.

In the NS data set, clean speech signals, recorded by 28 gender-balanced speakers, were subject to 10 different noises obtained from the DEMAND database [33] at 0 dB, 5 dB, 10 dB, and 15 dB SNRs. From the clean subset of this data set, we randomly selected 1,800 samples for the clean class, and 1,800 non-overlapping samples from the noisy subset for the noisy class.

In the NRS database, the noisy reverberant speech is created by convolving a clean signal with a RIR and adding it to a noisy signal that was also convolved with a room impulse response. Thus, we randomly selected 1800 samples for the noisy-reverberant class. To prepare data for the reverberant and nonlinear distortion classes, we selected two disjoint subsets of 1800 samples from the clean part of the data set, and degraded them in a similar way as for creating reverberant and nonlinear distortion classes for the pathological voices.

3.2 Acoustic features

We normalized the signals by subtracting the mean and dividing by the absolute maximum amplitude. Then, for the input to the CNNs, we segmented a signal into frames of 30 ms with 10 ms overlap using a Hamming window. Then, for each frame of a signal, we computed 40 channels log-mel spectrogram together with the first and second derivatives.

¹The babble, restaurant and street noise files were taken from <https://www.soundjay.com>, the office noise was taken from <https://freesound.org/people/DavidErbr/sounds/327497>, the white noise was taken from https://www.audiocheck.net/testtones_whitenoise.php, and the wind noise was taken from <https://www.iks.rwth-aachen.de/forschung/tools-downloads/databases/wind-noise-database>.

As the input to the GMM-UBM system, we used MFCCs computed by using a 30 ms Hamming window with 10 ms overlap and a 27 channel mel-scale filter bank. For each frame of a signal, 13 coefficients, including the log-energy of the frame, along with the first and second derivatives of the MFCCs have been calculated to form a 39-dimensional feature vector. We used the same values for the parameters of the baseline system as were used in [4] to reproduce their results.

3.3 Configuration parameters

All CNN networks in our experiments were trained 20 epochs by using SGD to minimize the binary-cross-entropy loss function defined in Eq. (4). The magnitude of the random fluctuations in the SGD dynamics is represented by the noise scale, ρ , which is proportional to the speed of convergence and defined as [34]:

$$\rho = \frac{\epsilon}{1-\nu} \left(\frac{N}{B} - 1 \right) \approx \frac{\epsilon N}{B(1-\nu)}, \quad (N \gg B), \quad (6)$$

where N is the number of training samples, ϵ is the learning rate, B is the batch size, and ν is the momentum of the SGD. In our experiments, the batch size in each epoch and the momentum of SGD were set to 64 and 0.9, respectively. We also exponentially decreased the learning rate from 0.01 to 0.0001 from one epoch to another.

For the baseline system, the number of mixture components is set to 1024 according to [4].

3.4 Performance metric

To evaluate the performance of the proposed system, we used the area under the receiver operating characteristic (ROC) curve (AUC). In the ROC curve, the true positive rate is plotted against the false positive rate for different decision thresholds of the scores. The AUC summarizes the ROC curve into a single number facilitating an easier comparison between different systems regardless of the decision threshold which is an application- and user-dependent parameter. The AUC value equals to 0.5 represents a chance level performance, while the AUC equals to 1 means a perfect separation of the classes.

4 Results and discussion

CNN is a complex, nonlinear transformer which can provide a rich variation of expressions of the input through the layers. By increasing the number of parameters, a better expression of the input can typically be achieved at the expense of increasing the risk of overfitting as the model can memorize specific details of the training data. Therefore, we first conduct an experiment to choose the optimal CNN architecture for the degradation detection problem and use it for the rest of the experiments. Then, after comparing the performance of the proposed method with the

baseline, we visually explain how the CNNs make decision for identifying a degradation in a speech signal.

In all experiments, we used 10-fold cross validation (CV) in which the samples were randomly divided into 10 non-overlapping and equal sized subsets. Then, 9 out of 10 subsets were used for training the models, and the remaining subset was used for evaluation. This procedure was repeated 10 times so that all subsets were used once for training and evaluating the model. It should be noted that for evaluation, we extended each test subset by adding 20 outlier samples, which do not contain relevant information with respect to the context of the data sets such as the bark of dog or a recording of whispered speech, to show whether the detectors can reject such outlier samples.

To investigate the best architecture for the CNN models, we compare the performance of CNN32, CNN64, CNN128, CNN256, and CNN512. These architectures are explained in Section 2.2 and illustrated in Fig. 1. In this experiment, we used the pathological voices. The results, reported in Table 1, show that the difference in performance between the various network architectures is marginal, particularly for noise detection in which all networks perform equally well. However, having a network of a simpler architecture which exhibits a higher performance is more desired to reduce the risk of overfitting. Considering the number of parameters of each model, mentioned in Section 2.2, and since the CNN128 outperforms others in identification of distortion and reverberation and has the most balanced complexity and accuracy for our application, we choose this architecture for the remaining experiments.

Once the optimal CNN architecture is selected, we can impartially compare the performance of the proposed system with the baseline. As explained in Section 2.3, the training phase in the baseline system consists of two steps, namely training a UBM with a large number of training samples from different degradation classes, and adapting degradation-dependent models with the corresponding training samples. For training the UBM, we used 8000 samples (1600 samples from each class). The remaining 1000 samples (200 samples from each class) were used for adapting and evaluating the degradation-dependent GMMs. To provide a fair comparison between the proposed method and the baseline system in terms of how

Table 1 Comparison between the performance of different CNN architectures on the pathological voice data set in the form mean AUC \pm 95% confidence interval

Detectors	CNN32	CNN64	CNN128	CNN256	CNN512
Noise	1.00 \pm 0.00	1.00 \pm 0.00	1.00 \pm 0.00	1.00 \pm 0.00	1.00 \pm 0.00
Distortion	0.98 \pm 0.01	0.98 \pm 0.01	0.99 \pm 0.00	0.98 \pm 0.01	0.98 \pm 0.01
Reverberation	0.90 \pm 0.01	0.91 \pm 0.01	0.93 \pm 0.01	0.91 \pm 0.01	0.89 \pm 0.01

The bold-faced numbers represent the best performance

Table 2 The performance of the CNN128 system when no parameters were shared across detectors and when the parameters in some layers were shared

Detectors	Independent training (first approach)	Sharing parameters (second approach)
Noise	1.00±0.00	1.00±0.00
Distortion	0.99±0.00	0.98±0.01
Reverberation	0.93±0.01	0.88±0.01

The results are in the form mean AUC±95% confidence interval

the data are used, we took two different approaches. In the first approach, we train each binary classifier from the scratch using all the corresponding training samples. In the second approach, on the other hand, we trained a multiclass classifier with the training samples used for training the UBM model. Then, using the samples exploited for adapting the degradation-dependent GMMs, we fine-tuned three binary classifiers from the trained multiclass classifier. In the fine-tuning step, we kept the parameters of the first and the second convolutional blocks frozen and adapted the parameters of the last convolutional block and the fully-connected layers. This way, similar to the baseline system, the parameters of the first two blocks were shared across each detector. Table 2 shows the performance of the CNN128 on the pathological voice data set when these two approaches were applied. We can observe that the models, particularly the reverberation detector, perform better when the classifiers were independently trained. Therefore, we used the first approach when comparing our proposed method with the baseline system.

Table 3 shows the performance of the baseline and the proposed systems. The results show that the proposed system outperforms the baseline for both pathological voices and running speech signals, particularly for identifying reverberation in pathological voices and additive noise in running speech. We can observe that both systems show a common tendency that the performance of the reverberation detector is much lower than the noise detector, mainly due to the false recognition of recordings in which noise and reverberation coexist, but the noise

Table 3 Comparison between the proposed method for degradation detection and the baseline system for pathological voice and normal running speech

Detectors	Pathological voice		Normal running speech	
	Baseline	Proposed	Baseline	Proposed
Noise	0.96±0.00	1.00±0.00	0.71±0.00	0.95±0.01
Distortion	0.90±0.01	0.99±0.00	0.83±0.01	1.00±0.00
Reverberation	0.75±0.00	0.93±0.01	0.84±0.01	0.99±0.00

The results are in the form mean AUC ±95% confidence interval, and the bold-faced numbers represent the best performance

is more dominant. Furthermore, the results indicate that the identification of reverberation in pathological voices is challenging for the baseline system. This is because unlike running speech, the temporal envelop of a sustained vowel is not peaky and, consequently, is not highly influenced by reverberation. Moreover, since the pitch contour in a sustained vowel remains almost the same over a short period of time compared to the running speech, the dynamic changes in the frequency domain are less influenced in sustained vowels than in running speech. These make the identification of reverberation more challenging for the baseline system. However, the CNN model can better distinguish these subtle differences.

On the other hand, since the frequency content and the characteristics of some types of background noises, such as babble, are similar to those of running speech signals, identifying additive noise in running speech is more challenging for the baseline system, while the CNN model could effectively detect the presence of the background noise in running speech. Given that for each noisy signal in the data set, we selected a random segment of a noise file and a random SNR value, and that the acoustic characteristics of the noise files used in these experiments vary in time (except for the white noise), the probability of observing noisy signals degraded by exactly the same noise segment with similar SNR value is very low. Therefore, based on the results, we expect the proposed system

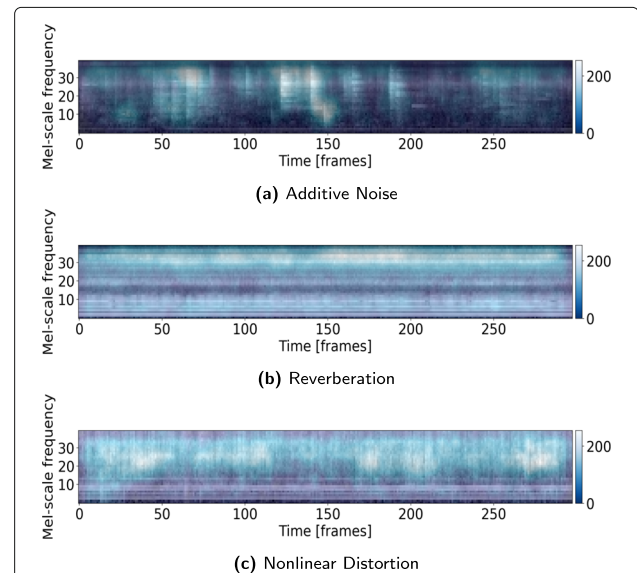
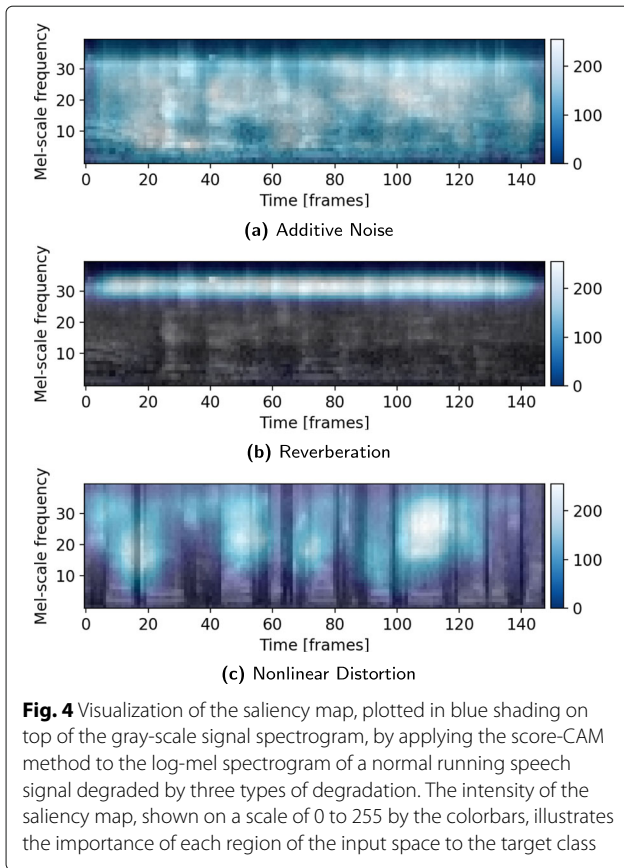


Fig. 3 Visualization of the saliency map, plotted in blue shading on top of the gray-scale signal spectrogram, by applying the score-CAM method to the log-mel spectrogram of a pathological voice signal (sustained vowel /a/) degraded by three types of degradation. The intensity of the saliency map, shown on a scale of 0 to 255 by the colorbars, illustrates the importance of each region of the input space to the target class



to be able to generalize for noise types not seen during the training phase.

Since deep learning models, such as CNNs, are associational machines which tend to learn the easiest path to associate the input data into the labels, one might suspect that a better performance of the proposed approach compared to the baseline system might be due to picking up spurious influences from some confounders in the data. Therefore, it is important to understand the basis on which the CNN models make a specific decision about the presence of degradation in a signal. There are a variety of techniques for understanding the behavior of complex deep learning models and how they make a particular decision [35]. Score-weighted class activation mapping (CAM) is one of these techniques which maps the internal representation in a CNN and provides a meaningful,

fine-grained visual explanation of complex CNN-based models [20].

In this method, different masks, referred to as the activation maps, are applied to the input image, which is the log-mel spectrogram of a speech signal in our experiments. Then, the prediction scores for each activation map is calculated and used as an indicator of the importance of that activation map. By overlaying the weighted activation maps on the input image, the parts of an image that are most influential to the score of the target class in prediction by the CNN model are highlighted. Figure 3 shows the saliency maps produced by applying the score-CAM method to a degraded pathological voice signal. We can observe the differences between the highlighted regions in the images depending on the type of degradation. For example, in Fig. 3a, a sustained vowel /a/ is degraded by a restaurant noise. It can be observed that the noise detector tends to focus on the wide-range areas in the log-mel spectrogram over the whole frequency, namely on both the patchy regions in the log-mel spectrogram, corresponding to clattering sounds of the tableware and plates, and some low-frequency regions, corresponding to the babble noise in the restaurant. On the other hand, as shown in Fig. 3b and c, the reverberation and distortion detectors tend to focus more on the continuous area along the temporal axis in the log-mel spectrogram and mainly in the high frequency regions. The results suggest that the high frequency regions are more influential and important in identification of distortion and reverberation. However, the tendency is a completely opposite in speaker recognition, which assigns great importance to the low frequency regions (about 200 Hz to 3 kHz) [36].

The saliency maps produced by applying the score-CAM method to a degraded normal running speech signal is shown in Fig. 4. We can observe that the tendency, i.e., that the noise detector focuses on the areas over the whole frequency and others focus on the high frequency region, is the same as the pathological voice. Interestingly, it can be seen that the distortion detector mainly focuses on the high frequency regions of the voiced frames (high power area) in the log-mel spectrogram. That is why, it is supposed that the nonlinear distortion appears remarkable when the original voice becomes loud.

Table 4 The impact of changing the lower and higher cutoff frequencies of the log-mel spectrogram on the performance of each detector for pathological voice signals

Detectors	f_{low} [Hz]				f_{high} [kHz]			
	0	300	700	2500	4.3	11	15	Nyquist
Noise	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
Distortion	0.99±0.00	0.98±0.01	0.98±0.01	0.98±0.01	0.97±0.01	0.98±0.01	0.97±0.01	0.99±0.00
Reverberation	0.93±0.01	0.93±0.01	0.91±0.01	0.86±0.02	0.83±0.02	0.88±0.01	0.92±0.01	0.93±0.01

The results are in the form mean AUC ±95% confidence interval

Table 5 The impact of changing the lower and higher cutoff frequencies of the log-mel spectrogram on the performance of each detector for normal running speech signals

Detectors	f_{low} [Hz]				f_{high} [kHz]			
	0	300	700	2500	4.3	11	15	Nyquist
Noise	0.95±0.01	0.92±0.01	0.94±0.00	0.91±0.01	0.99±0.00	1.00±0.00	0.99±0.00	0.95±0.01
Distortion	1.00±0.00	1.00±0.00	0.99±0.00	0.99±0.00	0.99±0.01	1.00±0.00	0.99±0.00	1.00±0.00
Reverberation	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.90±0.01	0.90±0.01	0.97±0.01	0.99±0.00

The results are in the form mean AUC ±95% confidence interval

To investigate further the importance of high frequency regions in degradation identification, we evaluated the performance of the proposed method using the log-mel spectrogram of different cutoff frequencies. The log-mel spectrogram is typically derived by applying triangular filters aligned at even intervals in mel-scale to the normalized STFT power. The linear frequency f in Hz can be converted to the mel-scale frequency m using the following equation [37]:

$$m = \Phi(f) = \frac{1000}{\ln(1 + 1000/700)} \ln\left(1 + \frac{f}{700}\right). \quad (7)$$

We define the low and high cutoff frequencies for the mel-scale filters as $m_{\text{low}} = \Phi(f_{\text{low}})$ and $m_{\text{high}} = \Phi(f_{\text{high}})$, respectively. The performance of each detector for pathological voice and normal running speech when changing the value of cutoff frequencies is reported in Tables 4 and 5, respectively. In these tables, the frequencies are shown in linear Hz-scale that can be converted to the mel-scale using the Eq. (7). It should be remarked that, in the sense of the Eq. (7), the amount of mel frequencies included in the frequency bands less than 300 Hz, 700 Hz, and 2.5 kHz is equivalent to those of included in the frequency bands of more than 15 kHz, 11 kHz, and 4.3 kHz, respectively. Despite of this fact, the performance of the reverberation detectors become significantly worse by decreasing the f_{high} to 11 kHz. However, increasing the f_{low} has only a limited impact on the performance of the reverberation detector. In contrast, the performance of the noise detector become slightly better by decreasing the f_{high} , probably due to the increase in the resolution of lower frequency regions. Meanwhile, the performance of distortion detectors stay pretty much the same even by changing the higher and lower cutoff frequencies. These results are well in consistent with the visual explanations in the previous experiments, and indicate that typical 8 kHz of sampling rate derived from telephone systems, is insufficient to identify the reverberation. We infer that a high frequency sound tends to be easily attenuated by a wall or other impediment objects in a room and, as a result, a damping appears in the high frequency region.

5 Conclusion

In this paper, we have proposed a new CNN-based approach for identifying degradation in speech signals. In this method, a set of CNN models, each responsible for detecting a particular type of degradation, has been used. The advantage of this method over the multiclass degradation detection methods is that parallel and independent detectors facilitate both detecting the presence of a combination of degradations in a speech signal and rejecting an outlier of a new type of degradation for which the models have not been trained. The CNNs were trained with the log-mel spectrogram of a large number of degraded speech signals. The experimental results using two different speech types, namely pathological sustained vowel sound and normal running speech show the effectiveness of the proposed approach in detecting degradations in signals which outperforms the state-of-the-art system. Furthermore, using the score-CAM technique, we visually explained how the CNN models make a specific decision in identifying degradation in signals. It also revealed that high frequency regions in log-mel spectrogram carry important information for identifying reverberation. It makes the identification of reverberation challenging when applying to telephone quality signals of 8 kHz sampling frequency.

Abbreviations

CNN: Convolutional neural network; RIR: Room impulse response; GMM: Gaussian mixture model; CAM: Class activation mapping; BN: Batch normalization; ReLU: Rectifier linear unit; STFT: Short-time fourier transform; SGD: Stochastic gradient descent; UBM: Universal background model; MFCC: Mel-frequency cepstral coefficient; MMPD: mPower mobile Parkinson's disease; PD: Parkinson's disease; SNR: Signal-to-noise ratio; dB: Decibel; Hz: Hertz; CELP: Code-excited linear prediction; NS: Noisy; NRS: Noisy-reverberant; ROC: Receiver operating characteristic; AUC: Area under the curve; CV: Cross validation.

Acknowledgements

Not applicable.

Authors' contributions

YS conducted the experiments. AHP and MGC designed the idea for the manuscript. All the authors contributed to the writing of this work. Moreover, all authors read and approved the final manuscript.

Funding

This work was funded by Independent Research Fund Denmark: DFF 4184-00056.

Availability of data and materials

The data set used for identification of degradation in pathological voice are available in the mPower Public Research Portal <https://www.synapse.org/#!Synapse:syn4993293/>. The data set used for identification of degradation in normal running speech are available in the Center for Speech Technology Research at University of Edinburgh <https://datashare.is.ed.ac.uk/handle/10283/2791> and <https://dx.doi.org/10.7488/ds/2139>. The database of room impulse responses is available here <http://www.ind.rwth-aachen.de/air>. The list of sampled data are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Received: 12 October 2020 Accepted: 15 January 2021

Published online: 05 February 2021

References

1. S. Ghai, R. Sinha, Adaptive feature truncation to address acoustic mismatch in automatic recognition of children's speech. *APSIPA Trans. Signal Inf. Process.* **5**, 1–13 (2016)
2. A. Alexander, F. Botti, D. Dessimoz, A. Drygajlo, The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications. *Forensic Sci. Int.* **146**, 95–99 (2004)
3. V. Mitra, A. Tsiartas, E. Shriberg, in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Noise and reverberation effects on depression detection from speech, (2016), pp. 5795–5799
4. A. H. Poorjam, M. S. Kavalekalam, L. Shi, J. P. Raykov, J. R. Jensen, M. A. Little, M. G. Christensen, Automatic quality control and enhancement for voice-based remote Parkinson's disease detection. *Speech Commun.* **127**, 1–16 (2021)
5. M. Fakhry, A. H. Poorjam, M. G. Christensen, in *European Signal Processing Conference (EUSIPCO)*. Speech enhancement by classification of noisy signals decomposed using NMF and Wiener filtering, (2018), pp. 16–20
6. J. H. L. Hansen, A. Kumar, P. Angkittrakul, Environment mismatch compensation using average eigenspace-based methods for robust speech recognition. *Int. J. Speech Technol.* **17**(4), 353–364 (2014)
7. B. W. Gillespie, H. S. Malvar, D. A. F. Florencio, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Speech dereverberation via maximum-kurtosis subband adaptive filtering, vol. 6, (2001), pp. 3701–3704
8. H. Kun, W. Yuxuan, W. DeLiang, S. W. William, M. Ivo, Z. Tao, Learning spectral mapping for speech dereverberation and denoising. *IEEE Trans. Audio Speech Lang. Process.* **23**(6), 982–992 (2015)
9. J. S. Abel, in *IEEE International Conference On Acoustics, Speech, and Signal Processing*. Restoring a clipped signal IEEE Computer Society, (1991), pp. 1745–1748
10. B. Defraene, N. Mansour, S. De Hertogh, T. Van Waterschoot, M. Diehl, M. Moonen, Declipping of audio signals using perceptual compressed sensing. *IEEE Trans. Audio Speech Lang. Process.* **21**(12), 2627–2637 (2013)
11. D. S. Williamson, D. Wang, Time-frequency masking in the complex domain for speech dereverberation and denoising. *IEEE/ACM Trans. Audio Speech Lang. Process.* **25**(7), 1492–1501 (2017)
12. T. Dietzen, S. Doclo, M. Moonen, T. van Waterschoot, Integrated sidelobe cancellation and linear prediction Kalman filter for joint multi-microphone speech dereverberation, interfering speech cancellation, and noise reduction. *IEEE/ACM Trans. Audio Speech Lang. Process.* **28**, 740–754 (2020)
13. I. Kodrasi, S. Doclo, Joint dereverberation and noise reduction based on acoustic multi-channel equalization. *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**(4), 680–693 (2016)
14. L. Ma, D. J. Smith, B. P. Milner, in *Eighth European Conference on Speech Communication and Technology*. Context awareness using environmental noise classification, (2003), pp. 1–4
15. J. M. Desmond, L. M. Collins, C. S. Throckmorton, Using channel-specific statistical models to detect reverberation in cochlear implant stimuli. *J. Acoust. Soc. Am.* **134**(2), 1112–1120 (2013)
16. S. V. Aleinik, M. Y. Nikolaevich, S. A. Vladimirovich, Detection of clipped fragments in acoustic signals. *J. Sci. Tech. Inf. Technol. Mech. Opt.* **92**(4), 91–97 (2014)
17. F. Bie, D. Wang, J. Wang, T. F. Zheng, Detection and reconstruction of clipped speech for speaker recognition. *Speech Commun.* **72**, 218–231 (2015)
18. A. H. Poorjam, J. R. Jensen, M. A. Little, M. G. Christensen, in *Proceedings of the Annual Conference of the International Speech Communication Association, InterSpeech*. Dominant distortion classification for pre-processing of vowels in remote biomedical voice analysis, (2017), pp. 289–293
19. A. H. Poorjam, M. A. Little, J. R. Jensen, M. G. Christensen, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A parametric approach for classification of distortions in pathological voices, (2018), pp. 286–290
20. H. Wang, M. Du, F. Yang, Z. Zhang, Score-CAM: improved visual explanations via score-weighted class activation mapping. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 111–119 (2020)
21. K. Simonyan, A. Zisserman, in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. Very deep convolutional networks for large-scale image recognition, (2015), pp. 1–14
22. J. Bjorck, C. Gomes, B. Selman, K. Q. Weinberger, in *Advances in Neural Information Processing Systems (NeurIPS)*. Understanding batch normalization, (2018), pp. 7694–7705
23. B. C. Moore, *An Introduction to the Psychology of Hearing*. (Emerald, Bingley, 2012)
24. K. Choi, G. Fazekas, M. Sandler, K. Cho, in *2018 26th European Signal Processing Conference (EUSIPCO)*. A comparison of audiosignal preprocessing methods for deep neural networks on musictagging, (Rome, 2018), pp. 1870–1874
25. B. M. Bot, C. Suver, E. C. Neto, M. Kellen, A. Klein, C. Bare, M. Doerr, A. Pratap, J. Wilbanks, E. R. Dorsey, S. H. Friend, A. D. Trister, The mPower study, Parkinson disease mobile data collected using ResearchKit. *Sci Data.* **3**, 1–9 (2016)
26. A. K. Ho, R. Iansek, C. Marigliani, J. L. Bradshaw, S. Gates, Speech impairment in a large sample of patients with Parkinson's disease. *Behav. Neurol.* **11**(3), 131–137 (1998)
27. I. R. Titze, D. W. Martin, Principles of voice production. *Acoust. Soc. Am.* **104**(3), 1148 (1998)
28. C. Valentini-Botinhao, Noisy speech database for training speech enhancement algorithms and tts models [online] (2017). Available: <http://dx.doi.org/10.7488/ds/2117>
29. M. Jeub, M. Schäfer, H. Krüger, C. Nelke, C. Beugeant, P. Vary, in *Proc. Int. Congress on Acoustics (ICA), Sydney, Australia*. Do we need dereverberation for hand-held telephony? (2010), pp. 1–7
30. M. R. Schroeder, New method of measuring reverberation time. *J. Acoust. Soc. Am.* **37**(6), 1187–1188 (1965)
31. M. R. Schroeder, B. S. Atal, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Code-excited linear prediction (CELP): high-quality speech at very low bit rates, vol. 10, (1985), pp. 937–940
32. C. Valentini-Botinhao, Noisy reverberant speech database for training speech enhancement algorithms and tts models [online] (2017). Available: <https://dx.doi.org/10.7488/ds/2139>
33. J. Thiemann, N. Ito, E. Vincent, The diverse environments multi-channel acoustic noise database: a database of multichannel environmental noise recordings. *J. Acoust. Soc. Am.* **133**(5), 3591–3591 (2013)
34. S. L. Smith, P. J. Kindermans, C. Ying, Q. V. Le, in *6th International Conference on Learning Representations*. Don't decay the learning rate, increase the batch size, (2018), pp. 1–11
35. M. Du, N. Liu, X. Hu, Techniques for interpretable machine learning. *Commun. ACM.* **63**(1), 68–77 (2019)
36. X. Lu, J. Dang, An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. *Speech Commun.* **50**(4), 312–322 (2008)
37. J. Harrington, S. Cassidy, *Techniques in speech acoustics*, vol. 8. (Springer Science & Business Media, Netherlands, 2012)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.