Digit Recognition Applied to Reconstructed Audio Signals Using Deep Learning

Anastasia-Sotiria Toufa and Constantine Kotropoulos
Department of Informatics, Aristotle University of Thessaloniki
Thessaloniki 54124, GREECE
Email: {anasttoufa,costas}@csd.auth.gr

***Abstract*—Compressed sensing allows signal reconstruction from a few measurements. This work proposes a complete pipeline for digit recognition applied to audio reconstructed signals. The reconstruction procedure exploits the assumption that the original signal lies in the range of a generator. A pretrained generator of a Generative Adversarial Network generates audio digits. A new method for reconstruction is proposed, using only the most active segment of the signal, i.e., the segment with the highest energy. The underlying assumption is that such segment offers a more compact representation, preserving the meaningful content of signal. Cases when the reconstruction produces noise, instead of digit, are treated as outliers. In order to detect and reject them, three unsupervised indicators are used, namely, the total energy of reconstructed signal, the predictions of an one-class Support Vector Machine, and the confidence of a pretrained classifier used for recognition. This classifier is based on neural networks architectures and is pretrained on original audio recordings, employing three input representations, i.e., raw audio, spectrogram, and gammatonegram. Experiments are conducted, analyzing both the quality of reconstruction and the performance of classifiers in digit recognition, demonstrating that the proposed method yields higher performance in both the quality of reconstruction and digit recognition accuracy.**

I. Introduction

Real world applications employ data, such as images, audio, video, or financial data, to solve tasks in a variety of domains. Although these data differ and have their own way of handling, they are signals, possessing common characteristics and sharing common principles. Many applications do not have access to the original information bearing signal due to memory, transmission, or privacy constraints. Instead, they have access to only a few measurements of the original signal and reconstruct the original signal based on these measurements prior to inference. Thus, it is necessary to develop efficient techniques to estimate the original signal. This process is known as signal reconstruction and its goal is the reconstructed signal to be as identical as possible to the original one.

Many methods for signal reconstruction are based on statistical signal analysis or extraction of features that allow efficient signal representation. One of the most common methods in signal reconstruction is compressed sensing. It exploits a favourable signal structure in a domain (e.g., sparsity) in order to solve an under-determined linear system [1]. Current techniques leverage the great performance of neural networks for signal reconstruction. In this direction, [2] is based on the principles of compressed sensing, assuming that the original signal lies in the range of a generator with fixed weights.

Representative examples of generators are variational autoencoders (VAEs) [3] and generative adversarial networks (GANs) [4], which have both the ability to model the distribution of data under examination. GANs consist of two models, namely the generator and the discriminator. The generator takes as input a low dimensional random noise vector z, and maps it to a higher dimensional space $G(z)$, generating new samples. The discriminator takes as input original and generated samples and classifies them in two classes, evaluating implicitly the quality of generated samples. The objective is the generated samples to be so realistic that the discriminator can not recognize them as fraud. The training between the two models is performed in an adversarial fashion, where at each iteration both networks improve their performance.

This work focuses on digit recognition applied to reconstructed audio signals, using deep learning techniques. The methodology exploits the findings in [2], extending them in the audio domain. A pretrained generator from a GAN is used that produces audio signals of digits [5]. The result of reconstruction is an audio signal that contains a digit with high probability. The processing of audio signals is a demanding task due to their nature. For example, speech signals have variability due to speaker pronunciation, speaker accent, and silence duration. Here, a new method for handling speech signals is proposed, demonstrating higher performance in both the quality of reconstruction and digit recognition accuracy. The key idea is to isolate the high energy segment of the signal, which most likely contains the digit, and discard the less active signal parts (e.g., silent segments). The identification of the most active segment of signal is performed through the detection of the segment with the highest energy. In this way, a more compact representation of the original signal is provided, preserving only the meaningful content. The measurement vector available for the reconstruction will be based on this high energy speech segment.

In the majority of cases, the reconstruction produces an audio digit. However, there are cases, where the reconstruction produces noise. The latter cases are handled as outliers. Because of the lack of labels in reconstructed samples, three unsupervised indicators are used for digit versus noise detection. The first indicator is the energy of reconstructed signal. The total energy is combined with the second indicator, the prediction of one-class Support Vector Machine (SVM) that operates as an outlier detector. Although, the correlation between these two indicators

is high, to obtain a more reliable detection, the confidence of digit recognition classifier is used as the third indicator. The underlying assumption is that a high confidence in classifier prediction refers to a non-noise (i.e., digit) signal that can easily be recognized. The combination of three indicators provides reliable detection of the reconstructed signals which contain digits and rejection of reconstructed signals which contain noise.

After the completion of digit versus noise detection, the proposed pipeline deals with digit recognition in reconstructed signals which most likely contain digits. Recognition is treated as multi-class classification, employing pretrained models on original audio recordings. Several classifiers are used, employing different representations of input data, such as raw audio, spectrograms, and gammatonegrams. The kind of input data and the choice of the extracted features affect model performance. The first classifier category comprises the k-nearest neighbor and nearest centroid, taking into consideration the distances among samples. Another category uses neural networks to perform classification by learning representative features from input data.

The contributions of this work are as follows:

- Technique [2] is extended into the audio domain.
- A novel method that uses only the highest energy segment of the reconstructed signal is proposed for audio reconstruction and digit classification. By confining ourselves to the highest energy segment of audio, better performance is observed in both the quality of reconstruction and digit recognition.
- A combination of three unsupervised indicators is used in order to detect digit vs noise for reconstructed signals.
- A pretrained digit classifier applied to raw audio signals outperforms classifiers trained on spectrograms and gammatonegrams, demonstrating a higher accuracy in digit recognition when reconstructed signals are employed.

The outline of the paper is as follows. Section II surveys related work. The proposed pipeline is described in Section III. Experimental findings are disclosed in Section IV and conclusion is drawn in Section V.

II. RELATED WORK

Compressed sensing (CS) [1] is a landmark in signal reconstruction. Extensive studies use compressed sensing in natural images [6] or video [7]. CS is also used in medical applications, such as computed tomography [8] and MRI images [9], [10]. CS is applied effectively in audio domain. It works in the context of sparse linear prediction [11], [12] or in building redundant dictionaries [13], while other works employ CS in the frequency domain [14].

Recent methods use deep learning for speech reconstruction. An inverse fast Fourier transform layer is proposed in [15]. Signals from the auditory cortex of patients with epilepsy are used for speech reconstruction in [16] and articulator movements are exploited in order to generate speech for people who have had laryngectomy in [17]. Audio-visual information is exploited for speech reconstruction through lip movements [18] or to generate natural-sounding acoustic speech signal [19].

Recently, GANs have attracted much attention in both generation of new, synthetic, samples and signal reconstruction. A combination of GANs and CS is proposed in [2], where sparsity is replaced with the requirement the input signal lies in the range of a pretrained generator. The optimization procedure focuses on determining a low dimensional vector in a latent space, which is used to generate the new sample. A similar formulation is adopted in [20] with the difference that the generator is involved in the training procedure. In the same direction, [21] provides an inverse mapping from the data space to a latent space. A general framework for training a single deep neural network that solves arbitrary linear inverse problems is established in [22].

To obtain an efficient audio and speech representation, most methods try to extract representative features, such as spectrograms [23], gammatonegrams [24], [25] or Mel frequency cepstral coefficients (MFCCs) [26]. Neural networks have shown a remarkable performance in speech recognition, using recurrent [27] or convolutional architectures [28]–[30]. Exploiting the ability of neural networks to model complex relations among data, many methods use raw audio signals for recognition [31], [32]

III. PROPOSED METHOD

The proposed method performs digit recognition in reconstructed audio recordings. The reconstruction procedure is positioned on the crossroads of CS and GANs, while the recognition procedure uses pretrained classifiers in original audio recordings. This work has been motivated by the research in [2], which has shown that GAN-based reconstruction is a variation of CS with effective performance. CS tries to estimate a sparse signal by solving an under-determined linear system with noisy linear measurements. The underlying assumption is the sparsity of the signal to be estimated in some domain. The variation in [2] replaces this assumption for signal structure by requiring that the signal to be estimated should lie in the range of a generative model.

A. Audio Compressed GAN

Let $x^* \in \mathbb{R}^n$ be the signal to be estimated, which is unknown during the reconstruction process. Let also y be available measurements, which are related to the signal to be estimated. The measurement vector used as input to the model, and is given by $y = Ax^*$, where $A \in \mathbb{R}^{m \times n}$ with $m \ll n$ is the measurement matrix, whose elements are random Gaussian numbers. That is, $y \in \mathbb{R}^m$, is a low dimensional representation of the signal to be estimated, comprising m measurements. Here, $n = 16384$ and $m = 500$. The reconstruction is performed through an iterative process, where at each iteration, the estimation $\hat{x}$ of the original signal is produced by a pretrained generator G of a GAN with fixed weights i.e.,

$$\hat{x} = G(z). \tag{1}$$

The generator of GAN takes as input a low dimensional noise vector $z \in R^k$, drawn from a predefined distribution p_Z, and

maps it to a high dimensional space $G(z) \in \mathbb{R}^n$ with $k < n$. The goal of optimization procedure is to find an optimal representation of z, and by extension of $G(z)$, in a way that the error between estimated measurement signal $\hat{y} = A\,G(z)$ and the available measurement vector y is minimized. The objective function is formally expressed in (2) and the optimization is performed through gradient descent:

$$\mathcal{L}(z) = ||AG(z) - y||_2^2. \tag{2}$$

Obviously, the quality of reconstruction depends on both the quality of generator and the signal to be estimated. On the one hand, the generator should be optimally trained, producing realistic samples $G(z)$, similar to those it has been trained on. On the other hand, the signal to be estimated x^* should lie on the range of GAN. For example, if a generator has been trained on ambient sounds of urban scenes, it would be impossible to provide a good reconstruction of speech signals.

This work uses the generator of the GAN proposed in [5], that is trained on audio recordings with digits from zero to nine. Audio signals require a detailed pre-processing, regarding their format and their sampling rate. For compatibility reasons, the same pre-processing as in [5] is applied. More specifically, regardless of the initial format, the signal to be estimated is transformed to float numbers and it is normalized by its maximum absolute value.

B. Extension to the Most Active Speech Segment

While CS is known for modeling adequately the domain it is applied to, here reconstruction from measurements collected from most active speech segment of original recordings $x \in \mathbb{R}^n$ is proposed. Such segment is identified via energy thresholding. Let $x_\Omega \in \mathbb{R}^{n'}$ be the most active speech segment with $n' < n$. Consequently, $y \in \mathbb{R}^m$ captures measurements from the most active speech segment and not the full audio recording, yielding a better reconstruction of the signal to be estimated x^*. The key idea of the proposed method is to focus only to the part of the signal with meaningful content, discarding the silent parts which do not provide any useful information.

To extract the aforementioned segment, speech activity detection is applied. That is, the energy of the audio samples, $s_i = |x_i|^2$, $i = 1, 2, \ldots, n$ is computed and the index $i_{\max}$ corresponding to maximum energy is determined. Then, x_Ω is the segment of length $n' = 6000$ around $i_{\max}$ extracted from x. Segment x_Ω is sampled through A and the measurements y are used as input for the reconstruction. The same loss function (2) is applied to estimate $\hat{x}$.

C. Digit Classification in Original Recordings

To perform digit recognition in reconstructed signals, pre-trained models in original audio recordings are required, demonstrating high classification performance. The underlying assumption is that if a classifier learns to recognize accurately digits by processing original recordings, then it may also classify reconstructed digits that may be distorted. Several classifiers have been tested, using different kinds of input. More specifically, k-nearest neighbors and nearest centroid, which work with distances between samples, as well as, a number of classifiers based on neural networks are used. The performance of k-nearest neighbor and nearest centroid classifier depends on descriptive representation of input data being able to separate different classes. For neural network-based classifiers, a suitable architecture has to be found, which can learn from input representations.

Three CNN-based classifiers have been examined. Each of them takes as input a different representation of audio signals. To begin with, raw audio recordings of length n have been used. The CNN applies 1D convolution and it consists of 5 layers with batch normalization and dropout. The other two classifiers are also CNNs, using as input spectrograms or gammatonegrams. Both representations are two-dimensional (2D) and can be treated as images. Spectrograms [23] offer a time-frequency distribution of the signal power, yielding a compact information of audio signals. Spectrogram has size 1025×33 and CNN architecture comprises 4 convolutional layers, followed by batch normalization and dropout. Gammatonegrams [24] are also 2D representations of audio signals, offering a time-frequency analysis matched to the human auditory system. Audio recordings are filtered by a gammatone filter-bank, which models the functionality of the mammalian cochlea in the ear. The difference between spectrograms and gammatonegrams is that spectrograms use a constant bandwidth across all frequency channels, while gammatonegrams have a wider bandwidth at higher frequencies than the lower ones in order to model ear's functionality. Gammatonegram has size 64×100. Hereafter, the three classifiers. that are based on neural networks, will referred to as raw audio classifier, spectrogram classifier, and gammatonegram classifier, respectively.

D. Noise vs Digit Detection

The goal of the proposed work is to initially reconstruct audio signals and then perform digit recognition in reconstructed signals. Through extensive experiments, audio compressed GAN is attested to be able to produce digits with high probability. However, there are cases where the reconstructed signal is either incomprehensible as digit or merely noise. Such ambiguous cases are considered as outliers and they should be detected and eliminated prior to digit recognition. This is a challenging problem due to the lack of labels for reconstructed signals, which makes difficult both noise versus digit detection and the evaluation of digit recognition, using reconstructed signal afterwards.

The procedure followed for digit versus noise detection is based on three unsupervised indicators. The first indicator is the total energy of the reconstructed signal. The underlying assumption is that noise samples have less energy than samples corresponding to clear digits. However, this indicator may not work effectively, if used in isolation. Thus an additional indicator should be tested. Taking into account that the reconstructed signal (1) corresponds to a digit in the most cases, instances when the reconstructed signal corresponds to noise could be treated as outliers. Such outliers can be detected by a one-class SVM, which provides a binary decision whether the

reconstructed signal is detected as noise or it is detected as digit. Alternatively, one may exploit uncertain estimation using deep ensembles [33]. There is high correlation between the energy of the signal and the prediction of one-class SVM, attesting that samples with low energy are considered as outliers. The combination of the two indicators makes noise versus detection more effective.

Despite the improvement in noise detection, there are still cases that one-class SVM incorrectly characterizes a reconstructed sample as noise or digit. In order to eliminate these cases, a third unsupervised indicator is used. The third indicator is the confidence of predictions obtained by digit recognition classifiers. More specifically, the output layer of classifiers (i.e., raw audio, spectrogram, gammatonegram classifier) is a softmax function, which produces a probability distribution associating the reconstructed audio signal to 10 digit classes. The predicted class is that corresponding to the highest probability. The probability of predicted class measures a confidence related to classifier decision, which is referred to as classifier confidence. Classifier confidence in used as the third indicator for noise vs digit detection. The underlying assumption is that if the prediction of a digit recognition classifier has low confidence, then the examined sample is noise. In order to combine effectively the three indicators, two thresholds are established for the noise vs digit detection, namely, *noise threshold* and *digit threshold*. For samples detected as noise by one-class SVM, if classifier confidence is greater than *noise threshold*, then the examined reconstructed audio is digit in reality. For reconstructed audio signals detected as digits by one-class SVM, if classifier confidence is less than a *digit threshold*, then the examined reconstructed signal is noise in reality.

In summary, three indicators are taken into consideration for detecting whether the input reconstructed audio signal is noise. The total energy of the signal, the prediction of outlier detector, and the digit classifier confidence. Their combination leads to a reliable procedure for noise versus digit detection, without being affected by the lack of labels for reconstructed signals.

E. Digit Recognition

As said previously, the ultimate goal of this work is digit recognition applied to reconstructed signals. The recognition is performed through pretrained classifiers in original, genuine, digit audio recordings. The key factor for efficient recognition is the quality of reconstructed signals. High quality reconstruction yields signals similar to original digit audio signals. If high reconstruction quality is combined with models offering high classification accuracy, then reconstructed signals can be easily recognized. To sum up, after reconstruction, it is attested whether the reconstructed signal (1) is digit or noise and if it is detected as digit, pretrained digit classifiers predict its class.

IV. Experiments

A. Dataset

The proposed model uses the *Speech Commands Dataset* [34]. The dataset consists of 65000 one-second long audio signals. Almost 2500 speakers have been recorded in uncontrolled recording conditions, ensuring high diversity in voices and accents. Here, a subset of the original dataset is used, comprising 18620 recordings associated to digits from zero to nine. The proposed audio compressed GAN uses the generator of WaveGan [5], which has also been trained on this subset, to provide reconstructions on the original signals. Digit recognition classifiers have been trained on 14896 original recordings and tested on 3724. To cope with the lack of labels for reconstructed signals, a test set of 150 reconstructed digit audio recordings that has been manually labeled is used.

B. Audio Compressed GAN

A set of experiments was conducted to select parameters, such as the number of measurements m, which denotes the size of vector y used in reconstruction, and the number of iterations needed for the best reconstruction. Other parameters refer to optimizer used and the learning rate of stochastic gradient descent. The performance of audio compressed GAN has been evaluated by the quality of reconstructions. $m = 500$ measurements are shown to be adequate for reconstructing original recordings of length $n = 16384$, with 1000 iterations. Measurement matrix A is the same in all reconstructions in order to have a common reference point. As for the training parameters, SGD optimizer has attained the best performance with fixed learning rate equal to 0.0001.

C. Digits Classification in Original Audio Recordings

To evaluate the performance of digit classifiers fed by original, genuine, digit audio recordings, accuracy and F_1 score have been computed. Five classifiers with different kinds of input have been tested. k-nearest neighbor, nearest centroid, and one CNN classifier employ raw audio signals as input. Another two CNNs use either the spectrogram or gammatonegram of audio signals as input, respectively. All classifiers, that are based on neural networks, have been trained using SGD optimizer with learning rate 0.01. The training has been completed with early stopping criteria.

Table I summarizes the performance metrics for all classifiers. The best results for k nearest classifier were obtained for $k = 10$ neighbors. Nearest centroid classifier reached top performance using the full dimension of original, genuine, digit audio recordings $n = 16383$. Dimensionality reduction by kernel Principal Component Analysis did not improve the performance of nearest centroid classifier. The inspection of the empirical findings in Table I reveals that CNNs yield the top performance. For the rest of the experimental analysis, only raw audio, spectrogram, and gammatonegram classifiers will be used.

D. Noise vs Digit Detection in Reconstructed Signals

After the reconstruction of audio signals, noise vs digit detection is performed. Because of the lack of labels in both tasks, the manually annotated dataset described in Section IV-A has been used in order to verify whether the three indicators and their combinations, discussed in Section III-D facilitate noise vs digit detection. The indicators are examined both independently

TABLE I
PERFORMANCE OF CLASSIFIERS APPLIED TO ORIGINAL DIGIT AUDIO RECORDINGS

	Accuracy		F_1 **score**
	Train Set	**Test Set**	**Test Set**
Raw Audio CNN	0.96	0.90	0.90
Spectrogram CNN	1	0.94	0.94
Gammatonegram CNN	0.99	0.97	0.97
10-Nearest Neighbor	0.21	0.11	0.10
Nearest Centroid	0.37	0.13	0.14

TABLE II
RATES OF TRUE NOISE SAMPLES IN TWO INTERVALS CREATED BY THRESHOLD IN CLASSIFIER CONFIDENCE

Confidence	**Raw Audio**	**Spectrogram**	**Gammatonegram**
Value	***Full Reconstructed Signals - 29 Noise Samples***		
$[0, 0.5]$	0.72	0.07	0.17
$(0.5, 1]$	0.28	0.93	0.83
	High Energy Segments - 45 Noise Samples		
$[0, 0.5]$	0.64	0.09	0.16
$(0.5, 1]$	0.36	0.91	0.84

and in combination, resulting in a statistical analysis for noise vs digit detection. The experimental analysis is two-fold, first it computes conditional probabilities given true noise signals and then statistical analysis uses only the detections delivered by the three indicators without having any knowledge about true noise samples.

When *full reconstructed signals* (i.e., reconstructed audio signals of size $n = 16368$) are employed, the true noise signals are 29 in total and the range of their energy is 174.83. When only *high energy segments* (i.e., the most active speech segment of length $n' = 6000$) are used, there are 45 reconstructed noise audio signals and their corresponding range of the energy is 105.78.

Taking into consideration the results of the second indicator, (i.e., one-class SVM outlier detection) for *full reconstructed signals*, out of 29 true noise signals, 24 are detected as noise and only 5 are detected as digits, yielding true negative rate and false positive rate 83% and 17%, respectively. Here, the results of one-class SVM show that noise signals are detected effectively in most cases. The combination of two indicators, namely the signal energy and one-class SVM outlier detection, shows that there is high correlation between them and attests our assumption that signals with low energy are noise signals. For *high energy segments*, out of 45 true noise signals, 21 are detected as noise and 24 as digits, yielding a true negative rate of 47% and a false positive rate of 53%, respectively. In this case, one-class SVM can not separate and detect efficiently noise samples.

Let us analyze the third indicator, i.e., classifier confidence. For each CNN classifier, a threshold of 0.5 is set for the classifier confidence in order to compute the true negative rates in each interval. The results are summarized in Table II. If raw audio classifier is employed, the majority of true noise samples have confidence in classifier prediction less than 0.5 for both *full reconstructed signals* and *high energy segments*. More specifically, the classifier confidence for true noise signals is less than 0.5 for 72% of *full reconstructed signals* and for 64% of *high energy segments*. These results attest our assumption that if digit recognition classifier confidence is low, here less than 0.5, the examined signals cannot be recognized, because they are noise. On the contrary, the confidence of spectrogram and gammatonegram classifier is very high in almost all true noise signals. In the case of *full reconstructed signals*, spectrogram classifier classifies 93% of true noise signals as digits with confidence greater than 0.5, while in the case of *high energy segments*, 91% of true noise signals are classified as digits with equally high confidence. These extreme high rates show that both spectrogram and gammatonegram classifiers can not find any difference between noise and digit signals and they classify noise signals incorrectly as digits with a high level of confidence.

The second part of statistical analysis does not use a priori knowledge of true noise samples. Instead, it uses only the results of the three indicators and their combinations. Combining the total energy of each reconstructed signal and the predictions of one-class SVM in *full reconstructed signals*, the range of energy in samples detected as noise by one-class SVM is 25.45, while the corresponding energy range for samples detected as digits is 360.62. In *high energy segments*, the corresponding ranges are 36.89 and 207.04. It is apparent that one-class SVM detects signals with low energy as noise and signals with high energy as digits. Thus, the predictions by one-class SVM are consistent with signal energy. A careful look at the predictions of one-class SVM for 150 *full reconstructed signals* reveals that 73 are detected as noise and 77 as digits, while for *high energy segments*, the detections are equal, i.e., 75 samples are detected as noise and another 75 samples are detected as digits.

The third indicator, the classifier confidence, shows high correlation with signals already characterized as noise by the other two indicators. To assess their correlation, the range [0,1] of classifier confidence that corresponds to the probability of digit recognition (i.e., classifier prediction) is split into ten equal intervals, representing histogram bins. For each bin, the total number of corresponding samples detected as noise by one-class SVM and classifier confidence lying in the specific interval is counted. The histograms regarding the number of signals detected as noise at each confidence interval for *full reconstructed signals* and *high energy segments* are plotted in Fig.1 and Fig.2, respectively.

As can be seen in Fig.1, in raw audio classifier, the majority of signals detected as noise appears into the confidence range from 0.4 to 0.7. On the contrary, most predictions of spectrogram and gammatonegram classifiers for signals detected as noise lie in the bin [0.9, 1]. The same study was performed for high energy segments. This observation verifies the assumption that low classifier confidence indicates noise samples due to

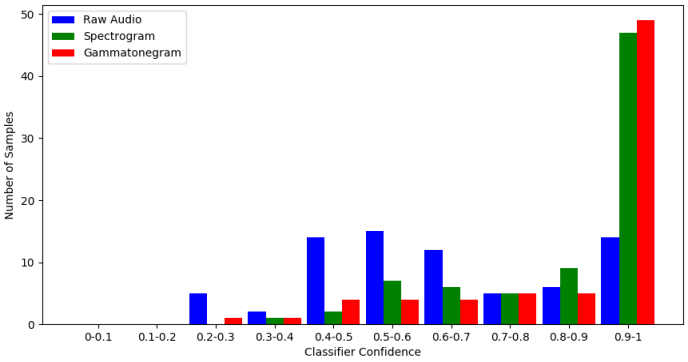

Fig. 1. Histogram of classifier confidence in ten intervals (i.e., bins) for signals detected as noise when *full reconstructed signals* are employed. Samples refer to the total number of signals lying in each bin.

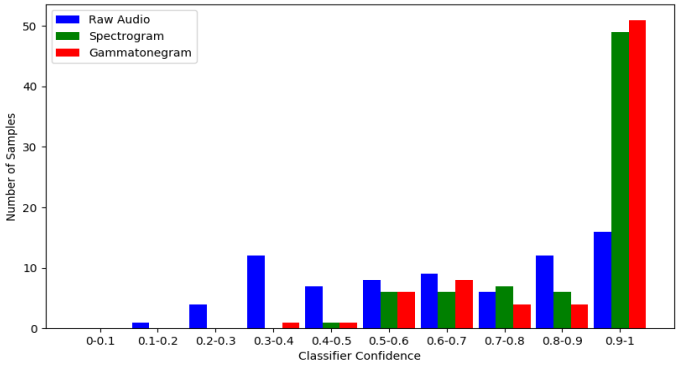

Fig. 2. Histogram of classifier confidence in ten intervals (i.e., bins) for signals detected as noise when *high energy segments* are employed. Samples refer to the total number of signals lying in each bin.

their ambiguous prediction. In Fig. 2, the histogram of signals detected as noise across all bins of confidence for raw audio classifier is evenly distributed across the ten intervals (i.e., bins). For spectrogram and gammatonegram classifiers same observations as in Fig. 1 are made.

E. Digits Recognition applied in Reconstructed Audio Data

In order to assess the combination of the three indicators in digit recognition, the two thresholds introduced in Section III-D are used. The *noise threshold* is used only for samples detected as noise by one-class SVM, while the *digit threshold* is used only for those samples detected as digit by one-class SVM. Taking into account the attested assumption that noise signals have low classifier confidence and digit signals have high classifier confidence, the goal is to find a suitable combination for the two thresholds in order to improve the performance in digit recognition. A higher value for *noise threshold* should be chosen, because signals detected as noise require high classifier confidence in order to be classified correctly. On the contrary, a lower value for *digit threshold* can be used, because signals detected as digits can be classified correctly more easily with low classifier confidence.

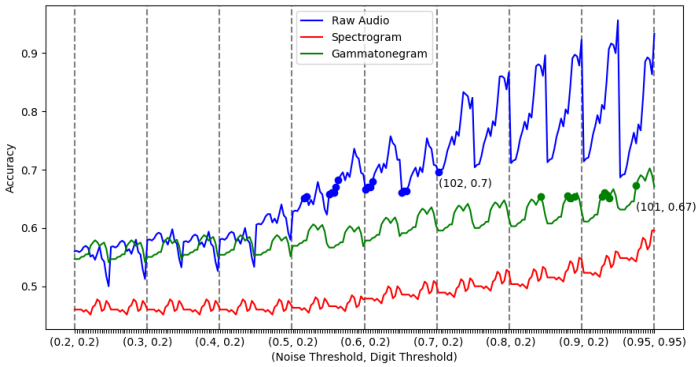

Fig. 3. Accuracy of vs. thresholds for *full reconstructed signals*. When recognition accuracy exceeds 0.65 and the total number of signals exceeds 100, both the total number of signals and accuracy are indicated in the plots.

All combinations for the two thresholds are tested from 0.2 to 0.9, using recognition accuracy as metric. If classifier confidence is lower than the threshold then the signal is discarded. Moreover, the total number of signals retained is also monitored. As the value of both thresholds increases, the accuracy increases too, but it is likely that too many signals are rejected and not taken into account in performance evaluation. There should be a trade-off between the number of signals retained and the recognition accuracy. That is, the goal is to find two thresholds that allow classifier predictions to be retained and yield the highest possible recognition accuracy.

In Fig.3 and Fig.4, values of the two thresholds are assessed with respect to recognition accuracy for each combination. In Fig.3, *full reconstructed signals* are examined. When accuracy exceeds 65% and the total number of signals is over 100, an operating point is overlaid on the plot. The best operating point is found for *noise threshold= 0.7* and *digit threshold= 0.2*, yielding *accuracy = 0.7* and *total number of signals= 102*, when raw audio classifier is used. Another competitive operating point is indicated, when gammatonegram classifier is employed.

The same experiment is repeated for *high energy segments* and the experimental findings are shown in Fig.4. Operating points are indicated, when accuracy exceeds 65% and the total number of signals exceeds 80. In this case, the total number of signals is smaller, in order to obtain more balanced results, because noise audio recordings are more in *high energy segments* than in full reconstructed signals. The best operating point is for *noise threshold = digit threshold = 0.65*, yielding *accuracy = 0.72* and total *number of signals = 80*, when raw audio is processed. A second competitive operating point is overlaid for gammatonegram classifier.

In order to measure the digit recognition accuracy, and indirectly assess the quality of reconstruction, it is very useful to exclude all true noise signals and confine ourselves to those audio recordings that are actually digits. As can be seen in Table III, the best performance is achieved by the raw audio classifier reconstructed signals obtained by *high energy segments*. Thus, when only the most active segment of audio recordings is

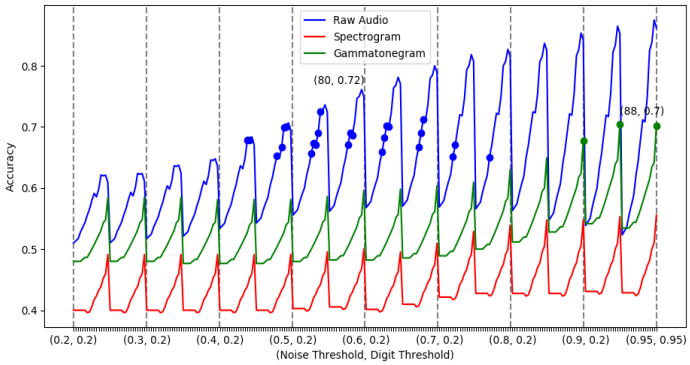

Fig. 4. Accuracy of vs. thresholds for *high energy segments*. When recognition accuracy exceeds 0.65 and the total number of signals exceeds 80, both the total number of signals and accuracy are indicated in the plots.

TABLE III
CLASSIFICATION ACCURACY AND F_1 SCORE WHEN TRUE NOISE SIGNALS ARE EXCLUDED

	Full Reconstructed Signals		
	Raw Audio	***Spectrogram***	***Gammatonegram***
Accuracy	**0.69**	0.57	0.68
F_1 score	0.51	0.45	0.56
	High Energy Segments		
Accuracy	**0.73**	0.57	0.69
F_1 score	0.58	0.51	0.67

exploited, the reconstructed signals have better quality and the digit recognition is more accurate.

The performance of the other two classifiers applied to spectrograms and gammatonegrams is inferior, leading to the conclusion that time-frequency representations may yield great performance in classification of genuine input audio signal, but they under perform when the signals to be recognized are reconstructed from a few measurements. On the contrary, when the amplitude of raw audio signal is processed, the discriminating information in original audio recordings is preserved in the reconstructed audio signals obtained from a few measurements. Despite any distortions in the reconstructed signals, the raw audio classifier can extract and exploit the discriminative features and achieve remarkable performance in digit recognition in both original and reconstructed signals.

V. CONCLUSION

A complete pipeline for digit recognition in reconstructed signals has been developed and assessed. A variation of compressed sensing technique has been used, where the original signal is constrained to lie in the range of a generator. The generator of a GAN has been used to generate reconstructed audio signals, corresponding to digits with high probability. A new method for reconstruction has been proposed where the most active segment of an original audio recording has been used in order to isolate the part of the signal, which contains meaningful information. For digit recognition, the first step is the detection of noise signals among the reconstructed audio signals that are considered as outliers. Signals that have been reconstructed efficiently and characterized as digits are classified by three pretrained classifiers trained on input raw audio, spectrograms, and gammatonegrams. The classifier trained on raw audio outperforms the classifiers applied to spectrograms and gammatonegrams when reconstructed signals are fed to the just mentioned classifiers. Through experiments, it has been demonstrated that the highest energy segment of the original signal provides reconstruction with better quality. The performance in recognition driven by such segment is shown to be higher, achieving a greater classification accuracy.

REFERENCES

[1] D. L. Donoho, "Compressed sensing," *IEEE Trans. Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[2] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," in *Proc. 34th Int. Conf. Machine Learning, JMLR -Vol 70*, 2017, pp. 537–546.

[3] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[5] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," *arXiv preprint arXiv:1802.04208*, 2018.

[6] L. Gan, "Block compressed sensing of natural images," in *Proc. 15th Int. Conf. Digital Signal Processing*, 2007, pp. 403–406.

[7] C. Chen, E. W. Tramel, and J. E. Fowler, "Compressed-sensing recovery of images and video using multihypothesis predictions," in *Proc. Conf. Signals, Systems and Computers*, 2011, pp. 1193–1198.

[8] G.-H. Chen, J. Tang, and S. Leng, "Prior image constrained compressed sensing (piccs): a method to accurately reconstruct dynamic ct images from highly undersampled projection data sets," *Medical Physics*, vol. 35, no. 2, pp. 660–663, 2008.

[9] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed sensing MRI," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 72–82, 2008.

[10] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MRI imaging," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.

[11] T. V. Sreenivas and W. B. Kleijn, "Compressive sensing for sparsely excited speech signals," in *Proc. 34th IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2009, pp. 4125–4128.

[12] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "Sparse linear predictors for speech processing," in *Proc. Ninth Annual Conf. Int. Speech Communication Association*, 2008, pp. 104–106.

[13] M. G. Christensen, J. Østergaard, and S. H. Jensen, "On compressed sensing and its application to speech and audio signals," in *Proc. 43rd Asilomar Conf. Signals, Systems and Computers*, 2009, pp. 356–360.

[14] P. Flandrin and P. Borgnat, "Time-frequency energy distributions meet compressed sensing," *IEEE Trans. Signal Processing*, vol. 58, no. 6, pp. 2974–2982, 2010.

[15] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. 40th IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2015, pp. 4390–4394.

[16] M. Yang, S. A. Sheth, C. A. Schevon, G. M. M. Ii, and N. Mesgarani, "Speech reconstruction from human auditory cortex with deep neural networks," in *Proc. 16th Annual Conf. Int. Speech Communication Association*, 2015, pp. 124–132.

[17] J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, "Direct speech reconstruction from articulatory sensor data by machine learning," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2362–2374, 2017.

[18] H. Akbari, H. Arora, L. Cao, and N. Mesgarani, "Lip2Audspec: Speech reconstruction from silent lip movements video," in *Proc. 43rd IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2018, pp. 2516–2520.

[19] A. Ephrat, T. Halperin, and S. Peleg, "Improved speech reconstruction from silent video," in *Proc. IEEE Int. Conf. Computer Vision Workshops*, 2017, pp. 455–462.

[20] A. Bora, E. Price, and A. G. Dimakis, "AmbientGAN: Generative models from lossy measurements." *in Proc. Int. Conf. Learning Representations*, vol. 2, p. 5, 2018.

[21] A. Creswell and A. A. Bharath, "Inverting the generator of a generative adversarial network," *IEEE Trans. Neural Networks and Learning Systems*, vol. 30, no. 7, pp. 1967–1974, 2018.

[22] J. Rick Chang, C.-L. Li, B. Poczos, B. Vijaya Kumar, and A. C. Sankaranarayanan, "One network to solve them all–solving linear inverse problems using deep projection models," in *Proc. IEEE Int. Conf. Computer Vision*, 2017, pp. 5888–5897.

[23] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[24] A. F. Pour, M. Asgari, and M. R. Hasanabadi, "Gammatonegram based speaker identification," in *Proc. 2014 Int. Conf. Computer and Knowledge Engineering*, 2014, pp. 52–55.

[25] P. Foggia, A. Saggese, N. Strisciuglio, and M. Vento, "Cascade classifiers trained on gammatonegrams for reliably detecting audio events," in *Proc. 11th IEEE Int. Conf. Advanced Video and Signal Based Surveillance*, 2014, pp. 50–55.

[26] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[27] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. 38th IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2013, pp. 6645–6649.

[28] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Proc. 2017 Int. Conf. Platform Technology and Service*, 2017, pp. 1–5.

[29] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Proc. 28th IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2003, pp. 8599–8603.

[30] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms." in *Proc. 18th Annual Conf. Int. Speech Communication Association*, 2017, pp. 1089–1093.

[31] D. Palaz, R. Collobert *et al.*, "Analysis of CNN-based speech recognition system using raw speech as input," IDIAP, Tech. Rep., 2015.

[32] J. Lee, T. Kim, J. Park, and J. Nam, "Raw waveform-based audio classification using sample-level CNN architectures," *arXiv preprint arXiv:1712.00866*, 2017.

[33] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017, pp. 6402–6413.

[34] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.