Digit Recognition Applied to Reconstructed Audio Signals Using Deep Learning

Anastasia-Sotiria Toufa and Constantine Kotropoulos

Dept. of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece

e-mail: costas@csd.auth.gr

ABSTRACT

Audio signal reconstruction and digit recognition in reconstructed signals are addressed.

- Compressed sensing is combined with generative adversarial networks (GANs).
- Pre-trained classifier on original audio recordings performs digit recognition, employing three input representations of signals.
- Most active speech segments perform better with respect to both quality of reconstruction and digit recognition accuracy.

Signal Reconstruction with Compressed Sensing and GANs

- Traditional compressed sensing method tries to estimate a sparse signal $\boldsymbol{x}^* \in \mathbb{R}^n$, using a set of available noisy measurements $\boldsymbol{y} \in \mathbb{R}^m$.
- Let $\boldsymbol{m} \ll \boldsymbol{n}$. The measurement vector is given by $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^*$, where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is the measurement matrix, whose elements are random Gaussian numbers.
- Alternatively, one assumes that the signal to be estimated lies in the range of a generative model [1].
- The estimated signal can be obtained by the generator $\boldsymbol{G}$ of WaveGan [2] with fixed weights: $\hat{\boldsymbol{x}} = \boldsymbol{G}(\boldsymbol{z})$.
- An optimal representation of $\boldsymbol{z}$ and consequently of $\boldsymbol{G}(\boldsymbol{z})$ is found, so that the loss function

$$\mathcal{L}(z) = || \underbrace{\boldsymbol{A}\,\boldsymbol{G}(\boldsymbol{z})}_{\hat{\boldsymbol{y}}} - \boldsymbol{y}||_2^2$$

is minimized by gradient descent.

Most Active Speech Segment Method

- Digit recordings from 0 to 9, 16384 samples long are used [3].
- To isolate the part of the signal that contains only the digit, the samples with high energy are detected.
- The energy of each audio sample is computed and the index corresponding to maximum energy $\boldsymbol{i_{max}}$ is determined.

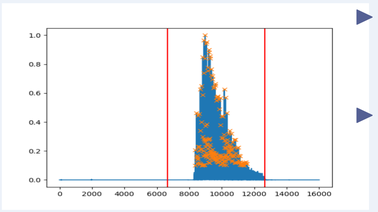

Figure 1: Isolation of the segment with digit through the detection of high energy samples.

- A segment 6000 samples long around $\boldsymbol{i_{max}}$ is extracted from initial signal $\boldsymbol{x}$.
- Vector $\boldsymbol{y}$ captures measurements only from this segment of the signal, yielding a better reconstruction of the signal to be estimated $\boldsymbol{x}^*$.

Noise vs Digit Detection

- Reconstruction failures produce signals being either incomprehensible as digits or merely noise. They are treated as outliers.
- Due to the lack of labels, the detection of such cases is performed using three unsupervised indicators:
  1. The total energy of reconstructed signal, assuming that noise signals have less energy than clear digits.
  2. The predictions of an one-class Support Vector Machine (SVM) providing a binary decision whether the reconstructed signal is detected as noise or digit.
  3. The confidence of predictions obtained by digit recognition classifiers.

Digit Recognition

- To perform digit recognition in reconstructed signals, the classifiers are trained in original audio recordings and they are tested in reconstructed signals.
- If a classifier learns to recognize accurately digits by processing original recordings, then it may also classify correctly the reconstructed digits that may be distorted. A variety of classifiers is used:
  1. $\boldsymbol{K}$-nearest neighbor and nearest centroid, using the distances among samples.
  2. Three CNN-based classifiers, using raw audio recordings, their spectrograms, or their gammatonegrams.

Experimental Evaluation

- Noise signal detection when true noise signals are known:

Table 1: Noise detection by thresholding CNN digit recognition confidence

Confidence Value	Raw Audio	Spectrogram	Gammatonegram
	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

  1. In full reconstructed signals, there is high correlation between signal energy and one-class SVM predictions, while in high energy segments, noise samples can not be detected effectively.
  2. Low classifier confidence corresponds to a noise signal (Table 1).

- When true noise signals are not known:
  1. One-class SVM uses signal energy to characterize the sample as digit or noise.
  2. Histograms of the number of signals detected by one-class SVM as noise, when classifier confidence is split into 10 bins are shown in Figures 2 and 3.

Experimental Evaluation (cont.)

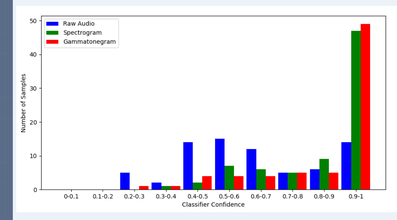

Figure 2: Histogram of noise samples for full reconstructed signals

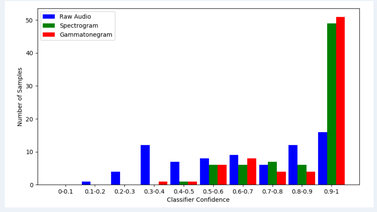

Figure 3: Histogram of noise samples for high energy segments

- Full reconstructed signals digit recognition accuracy $\mathbf{0.70}$; Total number of signals retained $\mathbf{102}$ (Figure 4).
- High energy segments digit recognition accuracy $\mathbf{0.72}$; Total number of signals retained $\mathbf{80}$ (Figure 5).

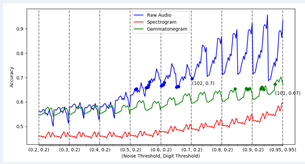

Figure 4: Digit recognition accuracy vs threshold combinations for full reconstructed signals

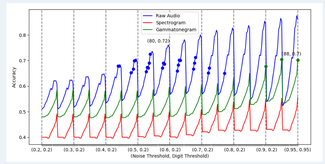

Figure 5: Digit recognition accuracy vs threshold combinations for high energy signals

Table 2: Digit recognition accuracy 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

- To indirectly assess the quality of reconstruction, digit recognition accuracy is measured when reconstructed signals are employed and all true noise samples are excluded (Table 2).

Conclusions

- Compressed sensing in combination with GANs can be extended effectively in audio domain.
- Most active speech segments ensure higher quality reconstruction and higher accuracy in digit recognition.
- Time-frequency representations 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.

References

[1] 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*, pp. 537–546, 2017.

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

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