Idiap Research Report

GRADIENT-BASED SPECTRAL VISUALIZATION OF CNNS USING RAW WAVEFORMS

Hannah Muckenhirn [a] Vinayak Abrol
Mathew Magimai.-Doss Sébastien Marcel

Idiap-RR-11-2018

JULY 2018

[a]Idiap Research Institute

Gradient-based spectral visualization of CNNs using raw waveforms

Hannah Muckenhirn,[1,2] Vinayak Abrol,[1] Mathew Magimai.-Doss,[1] Sébastien Marcel[1]

July 16, 2018

Abstract

Modeling directly raw waveform through neural networks for speech processing is gaining more and more attention. Despite its varied success, a question that remains is: what kind of information are such neural networks capturing or learning for different tasks from the speech signal? Such an insight is not only interesting for advancing those techniques but also for understanding better speech signal characteristics. This paper takes a step in that direction, where we develop a gradient based approach to estimate the relevance of each speech sample input on the output score. We show that analysis of the resulting "relevance signal" through conventional speech signal processing techniques can reveal the information modeled by the whole network. We demonstrate the potential of the proposed approach by analyzing raw waveform CNN-based phone recognition and speaker identification systems.

Index Terms: deep learning, CNN visualization, raw waveforms

1 Introduction

Traditionally automatic speech processing involves feature extraction followed by statistical modeling. The typical features being short-term spectral based features, which are extracted by applying speech production and speech perception knowledge. In recent years, with advances in neural networks especially deep learning, there is interest in reducing as much as possible hand crafted feature extraction. For instance,

1. by modeling intermediate representations such as filterbank outputs with a linear [1] or Mel scale [2] and spectrograms [3, 4]; or
2. by directly modeling raw speech signal [5, 6, 7, 8, 9, 10, 11, 12] using convolution neural networks (CNNs) at the input stage.

The interest of this paper lies in the latter case, where there is limited understanding about the information that is being modeled by the CNNs. Depending upon how the block processing is set or determined, we can split the approaches into two categories. First category, where the block processing is based on standard short-term or "segmental" processing (processing signal of about $1-3$ pitch period duration) [13, 6, 9, 14]. In the context of speech recognition, in [6] it was observed that the CNN filters modeling 35ms of speech signal tend to behave as a log-spaced frequency selective filter-bank. Whilst, in [7], some of the filters in the second convolution layer were found to behave like multi-resolution RASTA filters. Second category, where the block processing is determined during the training process in a task dependent manner [5, 11, 12]. In this case, it was found that for speech recognition the first layer of the CNN models "sub-segmental" speech signal (signal of duration below one pitch period) and captures formant information [15, 16]. In speaker recognition task, it was found that segmental modeling focuses on voice source related [12], while sub-segmental speech modeling focuses on vocal tract system related speaker discriminative information [17]. Similar observations have been made for the task of gender recognition [18]. These understandings are limited in the sense that they have been gained by analyzing the convolution layer(s). They not necessarily reveal the information that is being modeled as a whole from the input speech.

In computer vision research, it has been shown that gradient-based methods via relevance signal can help in visualizing the influence of each pixel in the input image on the prediction score [19, 20, 21, 22]. Inspired from that work, this present paper develops a gradient-based signal level and spectral level relevance map extraction approach to understand the task-dependent information modeled by the CNN-based system. In this approach, for a given input-target pair, the contribution of each input speech signal sample is first estimated and then analyzed using speech signal processing techniques. To the best of our knowledge, this is the first work which enables to visualize and analyze what is learned by an entire neural network trained on raw waveforms.

Section 2 presents relevant background work. Section 3 presents the proposed gradient-based visualization approach and Section 4 demonstrates its utility through phone recognition and speaker identification case studies. Finally, in Section 5 we conclude.

2 Relevant Background

In this work, we focus on the second category of CNN-based approach where block processing of the signal is often determined during the training process. Specifically, the approach that was first proposed for phone/speech recognition [5, 16] and has been later extended to other speech processing tasks, such as speaker recognition, presentation attack detection, gender recognition. As illustrated in Fig. 1, the network architecture consists of convolution layers followed by a multilayer perceptron. Each output o^c, $c = 1, \dots, C$ corresponds to a class. The parameters of the CNN and the MLP are jointly trained using cross entropy criterion.

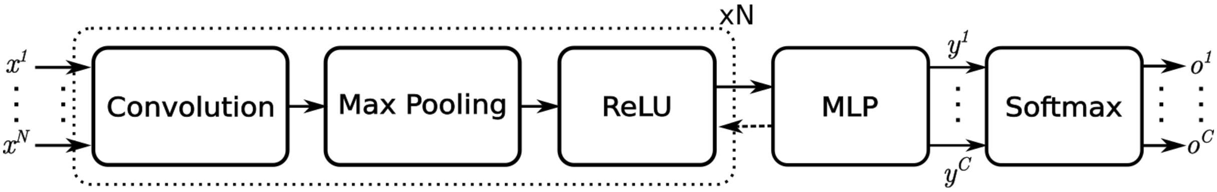


Figure 1: Architecture of the raw waveform based CNN system.

Fig. 2 illustrates the first convolution layer processing. At each time frame, the CNN takes as input a signal of length w_{seq}. This varies across applications. For instance, for speech recognition it is between 250-310 ms while for speaker recognition it is about 500 ms. kW and dW are the kernel width and kernel shift, respectively, which decides the block processing applied on the signal. n_f denotes the number of filters in the convolution layer.

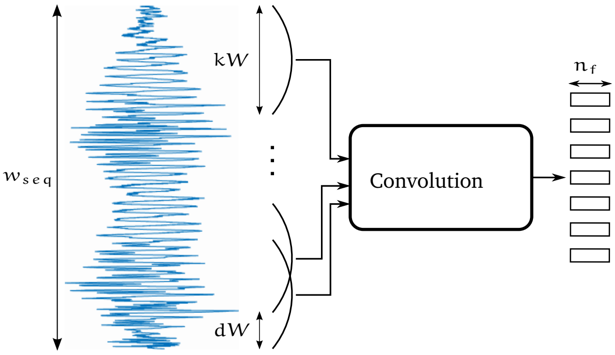


Figure 2: Illustration of first convolution layer processing.

In order to gain insight into the information that is being modeled, two level of analysis have been proposed [16]. First level of analysis is visualization of the cumulative frequency of the learner filters:

$$F_{cum} = \sum_{k=1}^{n_f} F_k / \|F_k\|_2, \tag{1}$$

where F_k is the frequency response of filter f_k. Second level of analysis interprets learned filters collectively as a spectral dictionary, leading to a sparsity point of view to understanding the network. This approach provides spectral information that is being modeled by CNN via analyzing the frequency response of filters to a given input. The magnitude frequency response $\mathbf{s}$ of the input signal $\mathbf{x} \in \mathbb{R}^{kW}$ is computed as:

$$\mathbf{s} = \left| \sum_{k=1}^{n_f} \langle \mathbf{x}, f_k \rangle \, \mathrm{DFT}[f_k] \right| . \tag{2}$$

If the atoms of the dictionary, i.e. $\{f_k\}$ are sines and cosines, then $\mathbf{s}$ would be the magnitude of the discrete Fourier transform of $\mathbf{x}$. In regular case, the dictionary is usually overcomplete and the inner product $\langle \mathbf{x}, f_k \rangle$ represents the weights (which are usually sparse) corresponding to the spectral contribution of atoms/filters.

3 Gradient-based Extraction of Spectral Relevance Map

In this section, we first motivate the need for extraction of spectral relevance map and then derive it.

3.1 Motivation

The analysis methods described in the previous section have helped in gaining insight into the works on speech recognition, presentation attack detection, speaker recognition and gender recognition [5, 11, 12, 18]. These analyses methods limits to the first convolution layer and provide insight into the spectral information being modeled by the first convolution layer. It does not however provides information about about what the CNN has learned as a whole or how the changes in the input affect the prediction score. To address this issue, we propose the use of gradient-based methods, where the gradient of a specific unit, which is usually the output unit that yields the highest score, is computed with respect to the input. The resultant gradient signal is referred to as “relevance” signal. This approach has helped in computer vision research to gain insight into the information in the image that the neural network focuses on. Several gradient-based methods have been proposed [21, 19, 20], and for CNNs they most differ by how the gradient of rectified linear units (ReLU) is computed during backpropagation. In this work we have considered the Guided backpropagation approach [20].

Fig. 3a shows an example of an input waveform fed to the CNN. The result of applying guided back-propagation given the input is shown in Fig. 3b. Note that it is not trivial to interpret the relevance signal as it is. Fig. 3c shows the auto-correlations of a short segment of input waveform and its corresponding time domain relevance signal. It can be observed that the time domain relevance signal contains information related to the periodicity of the speech signal. This suggests that spectral level interpretation could provide better insights.

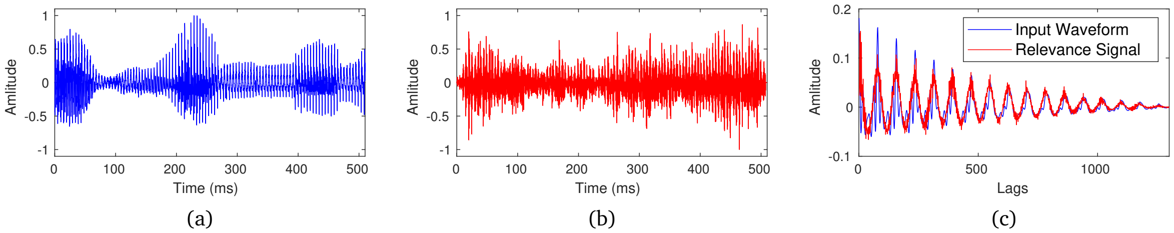


Figure 3: (a) Input Waveform, (b) Time domain relevance signal and (c) Autocorrelation plot.

3.2 Adaptation to frequency domain

Let $\mathbf{x} = [x_0 \dots x_{N-1}]$ is a raw audio frame ($N = w_{seg}$), belonging to class c, which is fed to a neural network. Next, discarding the softmax layer so as to remove influence from other classes, consider y^c the output unit corresponding to the class c. The gradient in the time domain with respect to input sample is defined as $f[n] = \frac{\partial y^c}{\partial x_n}$, $n = 0, \dots N-1$. Similarly, the gradient in the frequency domain can be expressed as $g[k] = \frac{\partial y^c}{\partial X_k}$ where $X_k = \sum_{n=0}^{N-1} x_n \exp(-i\frac{2\pi kn}{N})$. Applying the chain rule, one can express the two measures as:

$$
\begin{aligned}
\frac{\partial y^c}{\partial X_k} &= \sum_{n=0}^{N-1} \frac{\partial y^c}{\partial x_n} \frac{\partial x_n}{\partial X_k} \\
&= \frac{1}{N} \sum_{n=0}^{N-1} \frac{\partial y^c}{\partial x_n} \frac{\partial \sum_{j=0}^{N-1} X_j e^{i\frac{2\pi j n}{N}}}{\partial X_k} \\
&= \frac{1}{N} \sum_{n=0}^{N-1} \frac{\partial y^c}{\partial x_n} e^{i\frac{2\pi k n}{N}} \\
&= \frac{1}{N} \sum_{n=0}^{N-1} f[n] e^{i\frac{2\pi k n}{N}}
\end{aligned}
\tag{3}
$$

Thus,

$$
g[k] = \mathrm{DFT}^{-1}\{f[n]\}, \tag{4}
$$

it is complex and symmetric. The spectral relevance map can be visualized by plotting the amplitude of the first half of the signal, i.e. $|g[k]|$, for $k = 0, \ldots, \lceil \frac{N}{2} \rceil - 1$.

It is worth noting that the derived result is valid for any linear transformation, invertible with respect to $\mathbf{x}$. In other words, if $\mathbf{X} = M\mathbf{x}$, where the M is invertible, then $\frac{\partial y^c}{\partial \mathbf{X}} = M^{-1} \frac{\partial y^c}{\partial \mathbf{x}}$. Thus, other transforms could also be investigated.

4 Case studies: Phone classification and Speaker Identification

In this section, we present case studies on phone classification and speaker identification to demonstrate the utility of the proposed gradient-based spectral visualization approach. For the sake of simplicity, we carry out the investigations with CNNs trained on TIMIT corpus, as this allows contrasting between the two tasks. We used Keras-TensorFlow framework [23, 24] to train the neural networks and to perform guided backpropogation to obtain relevance signal. For analysis of the relevance signal, we use Praat toolkit [25] and MATLAB [26]. Section 4.1 presents the phone classification study and Section 4.2 present the speaker identification study.

4.1 Phone Classification

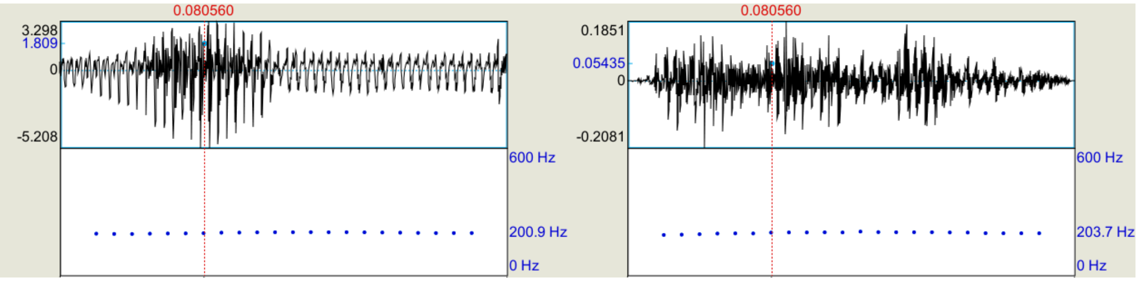
This section first presents the description of CNN-based phone classification system that is analyzed. We then present visualization of the relevance signal. Finally, a study quantifying the observations made in the visualization in an objective manner is presented.

4.1.1 System description

We trained a phone classifier on TIMIT dataset following the protocol that is used to benchmark phone recognition systems. We chose the hyper-parameters of the system with one hidden layer from the existing work in [16]. The hyper-parameters are presented in Table 1. In the original study the hyper-parameters were obtained through cross validation on the development set. The system yields phone error rate of 22.8% on the development set, and 23.6% on the test set.

Table 1: CNN architecture for phone classification. The input to the network is of length 250ms speech signal. n_f denotes the number of filters in the convolution layer. nhu denotes the number of hidden units in the hidden layer. kW denotes kernel width. dW denotes kernel shift (stride). Mpool+ReLU refers to max pooling followed by ReLU activation.

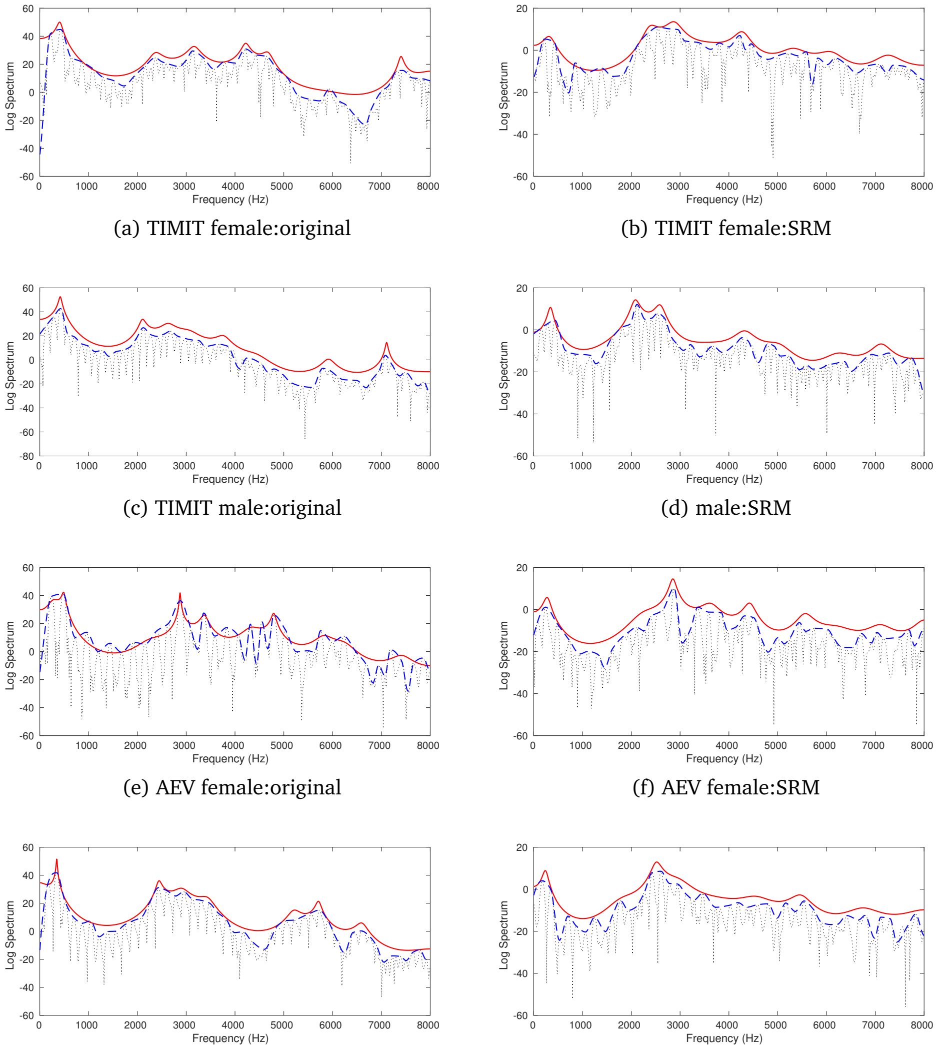
Layer	kW	dW	n_f/nhu
Conv1	30	10	80
Mpool+ReLU	3	3	-
Conv2	7	1	60
Mpool+ReLU	3	3	-
Conv3	7	1	60
Mpool+ReLU	3	3	-
MLP	-	-	1024



(a) Original

(b) Short Kernel

Figure 4: F0 contours for example waveform and corresponding relevance signal obtained for phoneme classification system.



(a) TIMIT female:original

(b) TIMIT female:SRM

(c) TIMIT male:original

(d) male:SRM

(e) AEV female:original

(f) AEV female:SRM

(g) AEV male:original

(h) AEV male:SRM

Figure 5: Example of original and spectral relevance maps (SRM) for vowel /iy/. (a-d): TIMIT; (e-h): AEV dataset, overlaid with spectral envelop (dashed:blue) and LP spectra (solid:red).

4.1.2 Visualization of relevance signal and spectral relevance map

Fig. 4 shows the original signal and relevance signal obtained for phone /ah/ uttered in TIMIT utterance 'sa1.wav' by speaker with ID 'fcjf0' along with the pitch frequency F0 contours for the two signals obtained using Praat. It can be observed at the signal level there are differences but the F0 contours are similar.

Figs. 5a-d show the short-term spectrum of the original signal and the short-term spectral relevance maps (SRMs) of /iy/ produced by a male speaker and a female speaker in exactly same phonetic context (i.e., speaking same text/word) in TIMIT corpus. Figs. 5e-h show the short-term spectrum of the original signal and the short-term spectral relevance maps (SRMs) of /iy/ produced by a male speaker and a female speaker in American English Vowels (AEV) dataset [27]. The analysis window size used was of length 25 ms. It can be observed that although the original signal and relevance signal differ in temporal domain, the harmonic structure and the envelop structure seem to be similar. Also, the network is able to generalize for unseen data (AEV in our case), and the formant structure of a particular sound unit is highlighted irrespective of the speaker.

4.1.3 Quantitative analysis

In order to ascertain that the relevance signal is indeed containing F0 and formant information, we performed a quantitative study on AEV dataset, since the steady state durations, F0 and formant information is available *a priori*. The analysis is done for 48 female and 45 male speakers following the standard protocol. In the steady state region, we computed F0 and first two formants (F1 and F2). The formants were computed using 16th order linear prediction analysis and is averaged over a context of 10 frames around the centre frame in the steady state region. We consider that the F0 and formant values are correct if it is with in the range F±Δ, where F is the F0 or F1 or F2 value and Δ is the respective standard deviation as specified in AEV dataset. Table 2 shows the average percentage accuracy of F0, F1 and F2 values for different phonemes. As it can be seen that the F0, F1 and F2 estimated from the relevance signal match well with the estimates provided in the AEV dataset. This shows that, despite the CNN modeling sub-segmental speech signal (about 2ms) at the input layer, the network as a whole is capturing both F0 and formant information.

Table 2: Average accuracy in (%) of fundamental frequencies, and formant frequencies of vowels produced by 45 male and 48 female speakers, estimated from relevance signal of AEV dataset.

		/ah/	/eh/	/iy/	/oa/	/uw/
F0	F	93	91	91	94	92
	M	92	90	89	93	90
F1	F	90	92	93	91	93
	M	88	92	92	89	93
F2	F	94	94	94	95	94
	M	94	93	94	94	93

4.2 Speaker Identification

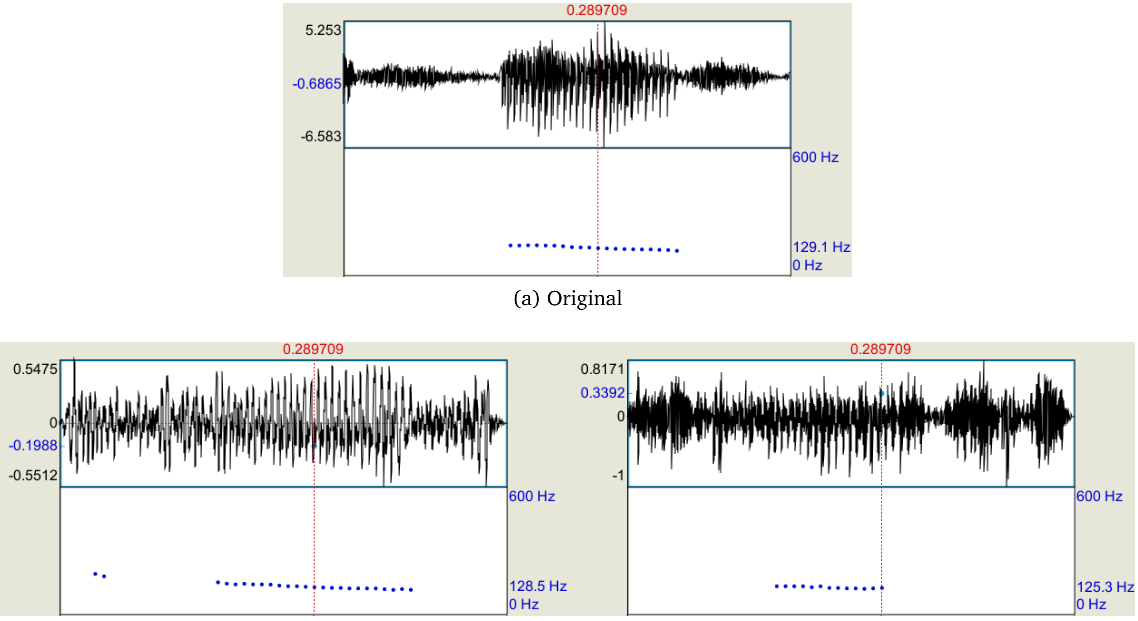
Section 4.2.1 presents the system description. Section 4.2.2 presents visualization of the relevance signal and Section 4.2.3 presents a quantitative analysis.

4.2.1 System description

In the speaker recognition approach proposed in [12], inspired from the speech recognition work, the hyper-parameters were determined based on cross-validation. It was found that the first convolution layer in this case models about 18ms speech signal and it captures voice sourced related information. Building on the observation that sub-segmental speech modeling in the case of speech recognition models formant information, the latter work [17] used the same first convolution layer configuration as speech recognition, with the aim being to ascertain whether that enables modeling vocal tract system related speaker discriminative information. We trained two CNN-based classifiers along those lines to classify the 462 speakers in the training set of the TIMIT phone recognition setup. We refer to these CNNs as segmental and sub-segmental CNNs. Table 3 provides the architecture information for the two cases. For each speaker, 9 utterances were used for training the CNN and 1 utterance is used for validation. The utterance-level accuracy obtained on the validation set is 98.3% (sub-segmental) and 94.5% (segmental), respectively.

Table 3: CNN architecture for speaker identification. The input to the network is of length 510ms. Definition of notations can be found in Table 1.

	Layer	kW	dW	n_f/nhu
sub-segmental	Conv1	30	10	80
	Mpool+ReLU	3	3	-
	Conv2	10	1	80
	Mpool+ReLU	3	3	-
	MLP	-	-	100
segmental	Conv1	300	10	100
	Mpool+ReLU	5	5	-
	Conv2	10	1	200
	Mpool+ReLU	5	5	-
	MLP	-	-	200



(a) Original

(b) segmental CNN

(c) sub-segmental CNN

Figure 6: F0 contours for example waveforms and corresponding relevance signal obtained for the two speaker identification systems.

4.2.2 Visualization of relevance signal and spectral relevance map

Fig. 6 presents an example speech signal and the relevance signal obtained for segmental CNN and sub-segmental CNN. Below each of the signal we also show F0 contours using Praat. It can be observed that the segmental CNN models F0 information better than the sub-segmental CNN.

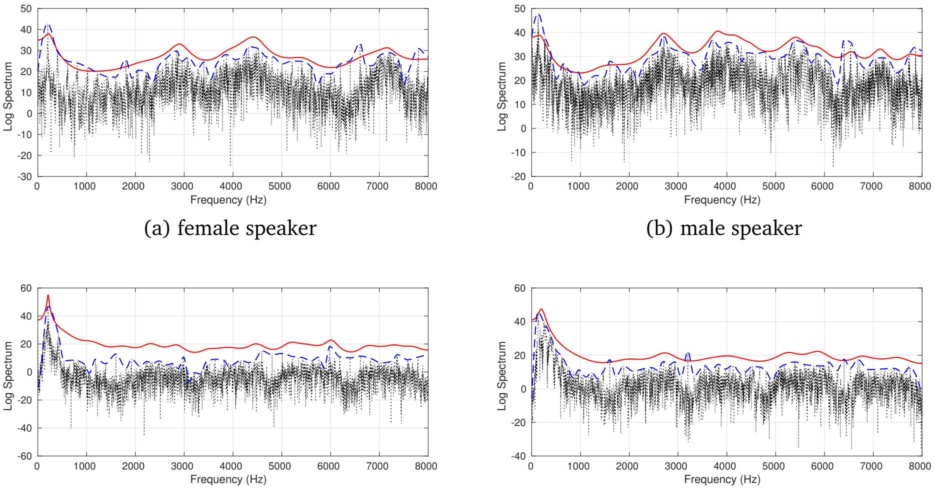
Figs. 7a-b shows the spectral relevance map on two waveforms belonging to two different speakers: one female and one male speaker, obtained for the sub-segmental CNN. The spectral relevance map is the averaged log spectra computed over a window of 25ms and shift of 10ms. As explained earlier, these two figures show which frequencies in the raw speech signal have a high influence on the prediction score. The observations on these two plots are consistent with what we found on many examples belonging to different speakers and are the following:

1. There is a highly localized peak, which appears to correspond to the value of the F0 of the speaker, as shown on the respective figures. For example, in Fig. 7a, this peak is at 190Hz. The F0 of the 10ms frame at the center of the input waveform estimated using Praat software is 202Hz. Similarly, in Fig. 7b, the second peak lies at 110Hz, while the estimated F0 is 114Hz.

2. There are two high frequency regions that are emphasized. A first region between 2000 and 3500 Hz and between 3500 and 5000 Hz. This is consistent with other studies [28, 29, 30], where authors performed an analysis of which frequency sub-bands are the most useful for speaker discrimination on the TIMIT database using either F-ratio measure [28, 29, 30] or vector ranking method [30]

They also found that mid/high frequencies were discriminative: respectively between 2500Hz and 4000Hz [28], between 2000Hz and 4000Hz [29] and between 3000Hz and 4500Hz [30].

Fig. 7c-d show the corresponding spectral relevance maps for segmental CNN. It can be observed that unlike the case of sub-segmental CNN, segmental CNN give more emphasis on the very low frequency bands, mainly around F0. A very localized peak can be observed around the fundamental frequency. These observations are in-line with the observations made in studies presented in [12].



(a) female speaker

(b) male speaker

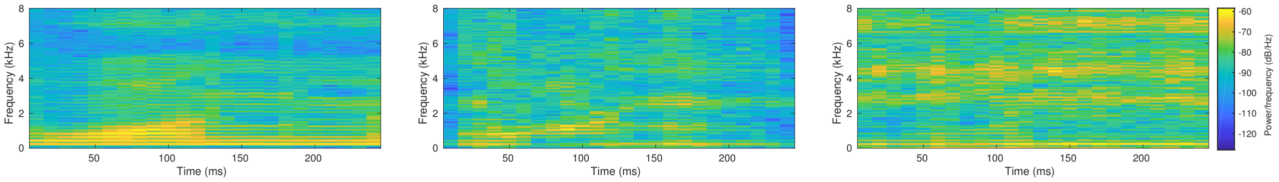
(c) female speaker

(d) male speaker

Figure 7: Example of average spectral relevance maps of two speakers: (a,b) CNN with short kernel and (c,d) CNN with long kernel, overlaid with spectral envelop (dashed:blue) and LP spectra (solid:red).

4.2.3 Quantitative analysis

In order to verify that the relevance signals contain F0 information, we conducted a quantitative study on TIMIT database by extracting and comparing the F0 contours of input speech waveform and with F0 contours of the relevance signal of segmental and sub-segmental CNN for all the ten utterances from 462 speakers. We performed the analysis only for the voiced frames in the original speech signal. The result is quantified in terms of the frame level F0 value deviation between F0 contour of the relevance signal with respect to the F0 contour of the input speech waveform. Approximately, 20% of the frames with F0 value zero in the F0 contour of the relevance signal are not considered in the calculation. For segmental CNN, the mean F0 deviation was 4Hz while for sub-segmental CNN it was 15Hz. This indicates that F0 information is captured by both segmental CNN and sub-segmental CNN. However, the sub-segmental CNN is selective, i.e. it seems to not model F0 information in all voiced frames.



(a) Original

(b) Phone Classification

(c) Speaker Identification

Figure 8: Spectrogram of an example waveform and corresponding spectral relevance maps obtained for phoneme classification CNN and speaker identification sub-segmental CNN.

4.2.4 Phone classification versus speaker identification

As mentioned earlier, sub-segmental CNN for speaker identification applies the same block processing as the phone classification CNN. In other words, both process 30 samples with a 10 samples shift. A question that arises: do the two systems focus on same kind of spectral information? Fig. 8 illustrates the difference in the information captured by the phone classification CNN and speaker verification sub-segmental CNN for /ah/ uttered by a TIMIT speaker. It can be observed that the phone classification CNN relevance signal retains well information related to the first two formants (around 1000 Hz) when compared to the speaker identification CNN relevance signal. We have performed informal listening tests on the relevance signals obtained with the two CNNs on a few TIMIT utterances. We have found that the relevance signal obtained with phone classification CNN is "intelligible", while the relevance signal of the speaker identification CNN is not. A detailed investigation along this line is part of our future work.

5 Discussion and Conclusion

Inspired from computer vision research, this paper proposed a gradient-based visualization approach for understanding the information modeled by CNN-based systems, which take raw signal as input. Through case studies on phone classification and speaker identification tasks, we showed that the relevance signal obtained through guided backpropagation can be analyzed using conventional speech signal processing techniques to gain insight into the information modeled by the whole neural network. These case studies also bring out the limitations of the spectral dictionary based approach to analyze first convolution layer (presented in Section 2). More precisely, spectral dictionary based analysis applied on phone classification task reveals that the CNN is modeling formant information [16] but it does not reveals that F0 information is also modeled. Similarly, on speaker identification task, a contrast between the findings of sub-segmental CNN analysis with the findings reported in [17] shows that F0 modeling and emphasis on high frequency regions is not revealed by the spectral dictionary based approach.

The relevance signal provides clues about the information modeled from the input signal by the whole neural network. However, it does not explains how the neural network is able to achieve that. Our future work will focus along that direction, where we aim to extend the proposed gradient-based approach to unravel the information modeled between the different intermediate layers and the output.

References

[1] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[2] Ossama Abdel-Hamid, Abdel-Rahman Mohamed, Hui Jiang, and Gerald Penn, “Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[3] Chunlei Zhang and Kazuhito Koishida, “End-to-end text-independent speaker verification with triplet loss on short utterances,” *Proc. of Interspeech*, 2017.

[4] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, “Voxceleb: a large-scale speaker identification dataset,” in *Proc. of Interspeech*, 2017.

[5] Dimitri Palaz, Ronan Collobert, and Mathew Magimai.-Doss, “Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks,” in *Proc. of Interspeech*, 2013.

[6] Tara N. Sainath, Ron J. Weiss, Andrew Senior, Kevin W. Wilson, and Oriol Vinyals, “Learning the speech front-end with raw waveform CLDNNs,” in *Proc. of Interspeech*, 2015.

[7] P. Golik, Z. Tüske, R. Schlüter, and H. Ney, “Convolutional neural networks for acoustic modeling of raw time signal in LVCSR,” in *Proceedings of Interspeech*, 2015, pp. 26–30.

[8] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn W. Schuller, and Stefanos Zafeiriou, “Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network,” in *Proc. of ICASSP*, 2016.

[9] Rubén Zazo, Tara N. Sainath, Gabor Simko, and Carolina Parada, “Feature learning with raw-waveform CLDNNs for voice activity detection,” in *Proc. of Interspeech*, 2016.

[10] Heinrich Dinkel, Nanxin Chen, Yanmin Qian, and Kai Yu, “End-to-end spoofing detection with raw waveform CLDNNS,” in *Proc. of ICASSP*, 2017.

[11] Hannah Muckenhirn, Mathew Magimai.-Doss, and Sébastien Marcel, “End-to-end convolutional neural network-based voice presentation attack detection,” in *Proc. of International Joint Conference on Biometrics*, 2017.

[12] Hannah Muckenhirn, Mathew Magimai.-Doss, and Sébastien Marcel, “Towards directly modeling raw speech signal for speaker verification using CNNs,” in *Proc. of ICASSP*, 2018.

[13] Zoltán Tüske, Pavel Golik, Ralf Schlüter, and Hermann Ney, “Acoustic modeling with deep neural networks using raw time signal for LVCSR,” in *Proc. of Interspeech*, 2014.

[14] Szu-Wei Fu, Yu Tsao, Xugang Lu, and Hisashi Kawai, “Raw waveform-based speech enhancement by fully convolutional networks,” in *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2017.

[15] Dimitri Palaz, Mathew Magimai.-Doss, and Ronan Collobert, “Analysis of CNN-based speech recognition system using raw speech as input,” in *Proc. of Interspeech*, 2015.

[16] Dimitri Palaz, Mathew Magimai.-Doss, and Ronan Collobert, “End-to-end acoustic modeling using convolutional neural networks for automatic speech recognition,” Idiap-RR Idiap-RR-18-2016, Idiap, 6 2016.

[17] Hannah Muckenhirn, Mathew Magimai.-Doss, and Sébastien Marcel, “On learning vocal tract system related speaker discriminative information from raw signal using CNNs,” in *Proc. of INTERSPEECH*, 2018.

[18] Selen Hande Kabil, Hannah Muckenhirn, and Mathew Magimai.-Doss, “On learning to identify genders from raw speech signal using CNNs,” in *Proceedings of Interspeech*, 2018.

[19] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. of International Conference on Learning Representations (ICLR)*, 2014.

[20] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *Proc. of International Conference on Learning Representations (ICLR)*, 2015.

[21] Matthew D. Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *Proc. of European Conference on Computer Vision*, 2014.

[22] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: removing noise by adding noise," in *ICML workshop on visualization for deep learning*, 2017.

[23] François Chollet et al., "Keras," 2015, Software available from https://github.com/fchollet/keras.

[24] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, Software available from tensorflow.org.

[25] Paul Boersma, "Praat, a system for doing phonetics by computer.," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[26] MATLAB, *version 9.4.0 (R2018a)*, The MathWorks Inc., Natick, Massachusetts, 2018.

[27] James Hillenbrand, Laura A. Getty, Michael J. Clark, and Kimberlee Wheeler, "Acoustic characteristics of american english vowels," *The Journal of the Acoustical society of America*, vol. 97, no. 5, pp. 3099–3111, 1995, http://homepages.wmich.edu/~hillenbr/voweldata.html.

[28] Tomi Kinnunen, "Spectral features for automatic text-independent speaker recognition," *Licentiates Thesis*, 2003.

[29] Laura Fernández Gallardo, Michael Wagner, and Sebastian Möller, "Spectral sub-band analysis of speaker verification employing narrowband and wideband speech," .

[30] Özgür Devrim Orman and Levent M Arslan, "Frequency analysis of speaker identification," in *2001: A Speaker Odyssey-The Speaker Recognition Workshop*, 2001.