Understanding and Visualizing Raw Waveform-based CNNs

Hannah Muckenhirn[1,2], *Vinayak Abrol*[3], *Mathew Magimai.-Doss*[1], *Sébastien Marcel*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland,
[3]Mathematical Institute, University of Oxford, UK

{hannah.muckenhirn,mathew,sebastien.marcel}@idiap.ch, abrol@maths.ox.ac.uk

Abstract

Modeling directly raw waveforms 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

Deep neural networks have become an integral part of many pattern recognition systems. In speech or audio related classification tasks, most deep learning systems are fed with intermediate features such as Mel-frequency cepstral coefficients (MFCCs) [1], filterbank outputs with a linear [2] or Mel scale [3] and spectrograms [4, 5]. However, in this case the input feature will have only limited spectral information constrained by the defined filter-bank type, magnitude compression or time-frequency resolution, which in turn influences the overall model architecture for a particular speech application. It has been shown that a perceptually designed filter bank is not always guaranteed to be the best for different speech applications. Hence, several studies have tried to address this issue with a waveform-based CNN that directly takes raw speech signal as input such as in speech recognition [6, 7, 8], emotion recognition [9], speaker recognition [10], voice activity detection [11], presentation attack detection [12, 13] and speech enhancement [14]. While these approaches have led to performance improvements, there is limited understanding about the information that is being modeled by the CNNs.

Depending upon whether the block processing is set or determined, we can split the approaches into two categories. In the first category, the block processing is based on standard short-term or "segmental" processing (processing a signal of about $1-3$ pitch period duration) [8, 11, 7, 14]. In the context of speech recognition, in [8] it was observed that the convolution filters, modeling 35ms of speech signal, tend to behave as a log-spaced frequency selective filter-bank. Whilst, in [15], some of the filters in the second convolution layer were found to behave like multi-resolution RASTA filters. In the second category, the block processing is determined during the training process in a task dependent manner [6, 13, 10]. 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 [16, 17]. In speaker recognition task, it was found that segmental modeling focuses on voice source related [10], while sub-segmental modeling focuses on vocal tract system related information [18]. Similar observations have been made for the task of gender recognition [19]. These understandings are limited in the sense that they have been gained by analyzing the first or second convolution layers. They do not necessarily reveal the information that is being modeled as a whole from the input speech.

On the other hand, understanding what information is modeled as a whole is an active field of research in computer vision. In particular, it has been shown that gradient-based methods can help in visualizing the influence of each pixel in the input image on the prediction score via a relevance map [20, 21, 22, 23]. While this visualization technique has been used in the case of neural networks fed with spectrograms [24], it is not straightforward to use it in the case of raw waveforms. This present paper develops a gradient-based temporal and spectral 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 sample is first estimated and then analyzed using signal processing techniques. To the best of our knowledge, this is the first work that enables to visualize and analyze what is learned by an entire neural network trained on raw waveforms.

Section 2 presents the relevant background work. Section 3 presents the gradient-based visualization approach and Section 4 demonstrates its utility through a phone recognition and a speaker identification case study. Finally, Section 5 concludes the paper.

2. Relevant Background

The gradient-based visualization technique described in this paper can be used on any type of neural network trained on raw waveforms. However, in this paper we focus on the approach that was first proposed for phone/speech recognition [6, 17] and has been later extended to other tasks, such as speaker recognition [10, 18], presentation attack detection [13] and gender recognition [19]. It corresponds to the second category of CNN-based approaches, where the block processing of the signal is determined during the training process. The network architecture consists of several convolution layers, followed by fully connected layers and a softmax layer.

This work was funded by the Swiss National Science Foundation (SNSF) through the project Unified Speech Processing Framework For Trustworthy Speaker Recognition (UniTS).

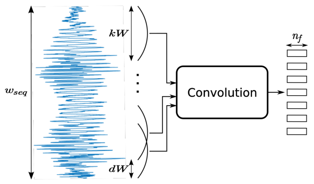


Figure 1: *Illustration of the first convolution layer processing.*

Fig. 1 illustrates processing at the first convolution layer. At each time frame, the CNN takes an input signal of length w_{seq}. kW and dW are the kernel width and shift, respectively, which decides the block processing applied on the signal. n_f denotes the number of filters in each layer. In order to gain insight into the information that is being modeled, two levels of analysis have been proposed [17]. The first level of analysis is the visualization of the cumulative frequency of the convolution filters:

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

where F_k is the magnitude spectrum of convolution filter f_k. The second level of analysis interprets the convolution filters collectively as a spectral dictionary, leading to a sparsity point of view to understanding the 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:

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

These analysis methods have helped in gaining insight into the works on speech recognition, presentation attack detection, speaker recognition and gender recognition [6, 13, 10, 19]. However, they are limited to the first layer and do not provide information about what the CNN has learned as a whole.

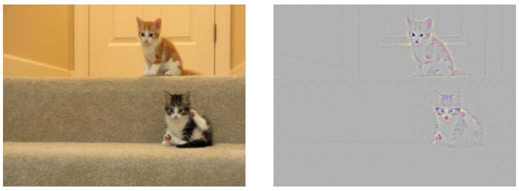
3. Gradient-based Visualization

In this section, we describe the gradient-based visualization method used in this paper.

3.1. Image processing

Visualization of what is captured by neural networks, especially by CNNs, is a very active field of research for image processing. Most visualization methods fall into three categories: 1) input perturbation-based methods, where the neural network is treated as a black box and the effect of altering the input image on the prediction score is measured, e.g., by occluding parts of the input [22]; 2) reconstruction-based methods [25, 20], where the idea is to synthesize or find among several images the input that maximizes the response of a unit of interest in the network; 3) gradient-based methods, which is the focuses of this paper.

In gradient-based methods, the gradient of a specific output unit, which is usually the one yielding the highest score, is computed with respect to each pixel of the input image. It measures how much a small variation of each pixel value will impact the prediction score. This corresponds to measuring the importance of each input value for the prediction. The result has the same size as the input image and is referred to as "relevance" map or "contribution" map. Several gradient-based methods have been proposed [22, 20, 21], and essentially they only differ by how the gradient of rectified linear units (ReLU) is computed during backpropagation. In this work, we use the Guided Backpropagation (GBP) method [21], as it has been shown to yield the sharpest results. In this method, the gradient at a ReLU layer is zero either if the gradient coming from above is negative or if the data value coming from below is negative. It is equivalent to computing the gradient of a ReLU function (as it is defined mathematically) but keeping only the gradients that have a positive values, i.e. a positive impact on the score prediction.

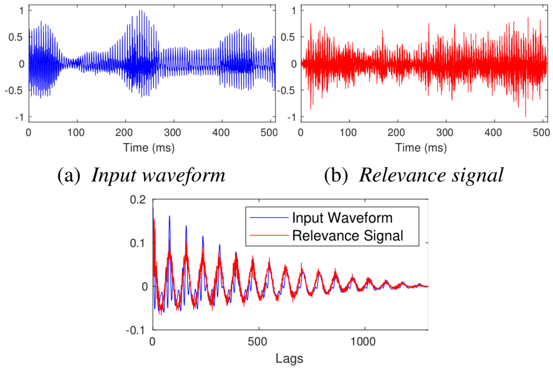


(a) *original image* (b) *relevance map*

Figure 2: *Original image, taken from the imageNet database and corresponding relevance map obtained with GBP method*

We show in Fig. 2 an example of such a visualization. The original image is taken from the imageNet database [26]. The relevance map, in Fig 2b, was obtained[1] with a VGG16 [27] trained on imageNet. We observe that the pixels that have a high impact on the classification results correspond to the two cats, while the pixels in the other parts of the image (stairs, wall, door...) are not important.

3.2. Speech processing



(a) *Input waveform* (b) *Relevance signal*

(c) *Autocorrelation*

Figure 3: *Analysis of the relevance signal obtained with GBP method with a CNN trained for speaker identification.*

The result of directly applying the Guided Backpropagation method in the case of raw waveforms is shown in Fig. 3a and 3b. Visualization in the time domain does not bring much insights into what important characteristics are extracted by the network because the results are difficult to interpret, as we do not have any visual cues as in the case of images. Fig. 3c shows the auto-correlations of a short segment of the input waveform and its corresponding relevance signal. It can be observed that the relevance signal contains information related to the periodicity of the speech signal. This suggests that spectral level interpretation could provide better insights.

[1] https://github.com/ramprs/grad-cam.

Let $\mathbf{x} = [x_0 \dots x_{N-1}]$ be a raw audio frame, 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 $f[n] = \frac{\partial y^c}{\partial x_n}$, $n = 0, \dots N-1$. We want to compute the gradient of the output unit y^c with respect to each frequency bin of the Fourier transform of the input waveform. That is, we want to visualize the impact of each frequency bin on the output. Thus, we want to compute $g[k] = \frac{\partial y^c}{\partial X_k}$ where $X_k = \sum_{n=0}^{N-1} x_n \exp(-i\frac{2\pi kn}{N})$. However, a real-valued non-constant function with complex-valued parameters does not fulfill the Cauchy-Riemann equations and is thus not differentiable. One can instead use the Wirtinger derivatives [28]. Applying the chain rule, one can express the two measures as:

$$\frac{\partial y^c}{\partial X_k} = \sum_{n=0}^{N-1} \frac{\partial y^c}{\partial x_n} \frac{\partial x_n}{\partial X_k} = \frac{1}{N} \sum_{n=0}^{N-1} \frac{\partial y^c}{\partial x_n} \frac{\partial \sum_{j=0}^{N-1} X_j e^{i\frac{2\pi jn}{N}}}{\partial X_k}$$
$$= \frac{1}{N} \sum_{n=0}^{N-1} \frac{\partial y^c}{\partial x_n} e^{i\frac{2\pi kn}{N}} = \frac{1}{N} \sum_{n=0}^{N-1} f[n] e^{i\frac{2\pi kn}{N}} \qquad (3)$$

Thus,

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

which is complex and symmetric. The spectral relevance map can be visualized by plotting $|g[k]|$, for $k = 0, \dots, \lceil\frac{N}{2}\rceil - 1$. The derivation is simplified by dropping the complex conjugate part in the Wirtinger chain rule and by assuming that $\mathbf{x}$ and its DFT have the same dimension N. For a more rigorous derivation, the reader is referred to [29].

While this approach is correct, the input signal $\mathbf{x}$ usually spans $250 - 500$ ms and cannot be assumed to be stationary. Thus, instead of computing the inverse DFT of $f[n]$ in (4) we compute the inverse short time Fourier transform.

4. Case studies: Phone classification and Speaker Identification

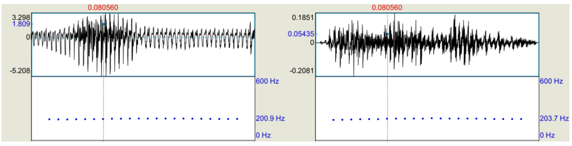
This section presents case studies on phone classification and speaker identification to show the utility of analyzing spectral relevance signals in understanding raw waveform CNNs.

4.1. Systems description

The phone classification and speaker identification systems are both trained on the TIMIT database. The phone classifier is trained following the protocol in [17]. The hyper-parameters are presented in Table 1. The input to the network is of length 250ms. The CNN is composed of three convolutional layers, followed by a fully connected layer. Each convolution is followed by a max pooling with a kernel width and shift of 3 samples and by a ReLU activation function. 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. The architecture and hyper-parameters of the speaker identification system are taken from [18] and detailed in Table 1. The architecture consists of two convolutional layers, followed by a fully connected layer. The CNN was trained to classify the 462 speakers in the training set of the TIMIT phone recognition setup. 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%.

Table 1: *Hyper-parameters of the phone classification and speaker identification systems. n_f denotes the number of filters in the convolution layer. n_{hu} denotes the number of hidden units in the fully connected (FC) layer. kW and dW denote kernel width and kernel shift (stride).*

	phone classification			speaker identification		
	kW	dW	n_f/n_{hu}	kW	dW	n_f/n_{hu}
Conv1	30	10	80	30	10	80
Conv2	7	1	60	10	1	80
Conv3	7	1	60	-	-	-
FC	-	-	1024	-	-	100



(a) *Original* (b) *Relevance signal*

Figure 4: *F0 contours for an example waveform and its relevance signal obtained for the phone classification system.*

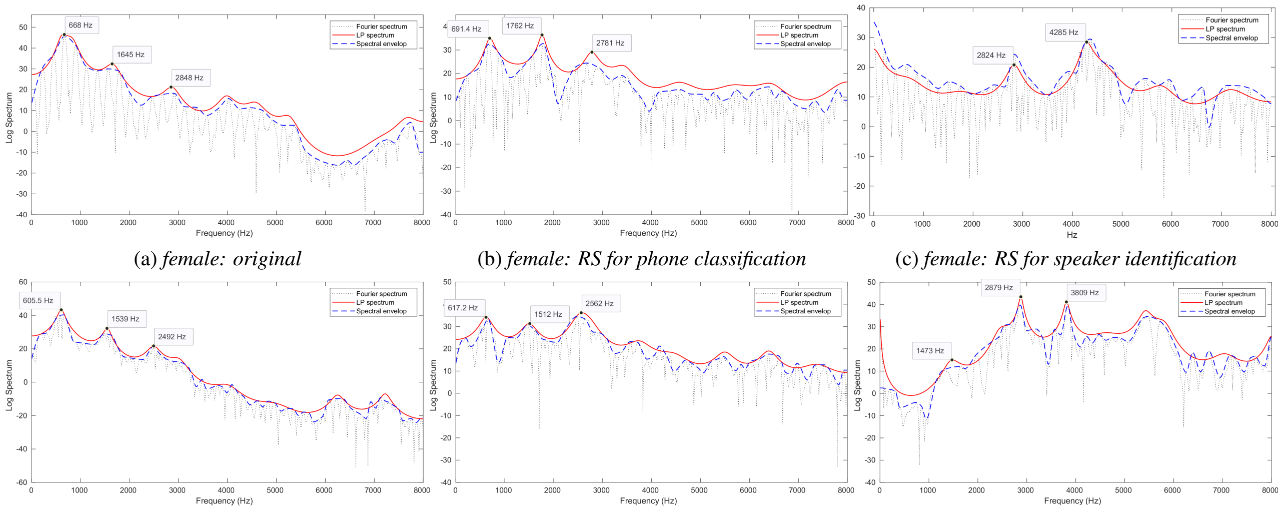
4.2. Phone Classification

4.2.1. *Visualization and analysis of relevance signals*

Fig. 4 shows the original waveform and the relevance signal corresponding to the phone /ah/ along with their pitch frequency F0 contours obtained using Praat toolkit [30]. We observe that the two signals are different in the temporal domain, however the F0 contours are similar. Fig. 5a and 5d show the short-term spectrum of the sound /ah/ produced by a male and a female speaker in exactly the same phonetic context (i.e., speaking the same text) in the TIMIT corpus. Fig. 5b and 5e show the short-term spectrum of the corresponding spectral relevance signals. The analysis window size used was of length 25 ms. We observed that, although the original signal and relevance signal differ in temporal domain, the harmonic structure and the envelop structure are similar. In particular the first and second formants.

4.2.2. *Quantitative analysis*

In order to ascertain that the relevance signal contains indeed F0 and formant information, we performed a quantitative study on the American English Vowels (AEV) dataset [31]. We chose this database because the steady state durations, F0 and formant information are available. 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 within the range F$\pm\Delta$, where F is the F0, F1 or F2 value and Δ is the respective standard deviation as specified in the 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.



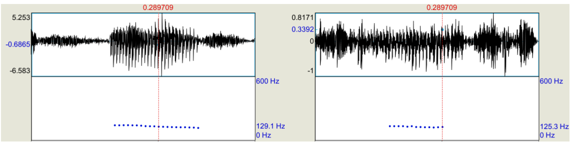
(a) *female: original* (b) *female: RS for phone classification* (c) *female: RS for speaker identification*

(d) *male: original* (e) *male: RS for phone classification* (f) *male: RS for speaker identification*

Figure 5: *Example of original and relevance signals (RS) for vowel* $/ah/$*, overlaid with spectral envelop and LP spectra.*

Table 2: *Average accuracy in (%) of fundamental 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



(a) *Original* (b) *Relevance signal*

Figure 6: *F0 contours for an example waveforms and its relevance signal obtained for the speaker identification system.*

4.3. Speaker Identification

Fig. 6 presents an example speech signal and the corresponding relevance signal for this system along with their F0 contours. similar to the case of phone classification, the two signals are very different in the time domain, however the F0 contours are similar. Figs. 5c and f show the short-term spectrum of the relevance signals, corresponding to the signals in Figs. 5a and d. The observations on these two plots are consistent with what we found on many examples belonging to different speakers and are the following. First, there is a peak in the low frequencies. Secondly, there are two high frequency regions that are emphasized: between 2000 and 3500 Hz and between 3500 and 5000 Hz. This is consistent with other studies [32, 33, 34] on TIMIT, where authors performed an analysis of which frequency subbands are the most useful for speaker discrimination using either F-ratio measure [32, 33, 34] or vector ranking method [34]. They found that mid/high frequencies were discriminative: respectively between 2500Hz and 4000Hz [32], between 2000Hz and 4000Hz [33] and between 3000Hz and 4500Hz [34].

The CNNs trained for speaker identification and for phone classification applies the same block processing on the raw waveforms, i.e, both process 30 samples with a 10 samples shift. However, we observe that the relevance signals are very different. We performed informal listening tests on the relevance signals obtained with the two CNNs and found that the relevance signal obtained with phone classification CNN is “intelligible”, while the relevance signal of the speaker identification CNN is not.

5. Discussion and Conclusion

Inspired from computer vision research, this paper extends the gradient-based visualization approach for understanding CNN-based systems, which take the 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 [17] 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 [18] 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.

6. References

[1] A.-R. Mohamed, G. E. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[2] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. 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.

[3] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, and G. 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.

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

[5] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: a large-scale speaker identification dataset,” in *Proc. of Interspeech*, 2017.

[6] D. Palaz, R. Collobert, and M. Magimai.-Doss, “Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks,” in *Proc. of Interspeech*, 2013.

[7] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, “Acoustic modeling with deep neural networks using raw time signal for LVCSR,” in *Proc. of Interspeech*, 2014.

[8] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, “Learning the speech front-end with raw waveform CLDNNs,” in *Proc. of Interspeech*, 2015.

[9] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, “Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network,” in *Proc. of ICASSP*, 2016.

[10] H. Muckenhirn, M. Magimai.-Doss, and S. Marcel, “Towards directly modeling raw speech signal for speaker verification using CNNs,” in *Proc. of ICASSP*, 2018.

[11] R. Zazo, T. N. Sainath, G. Simko, and C. Parada, “Feature learning with raw-waveform CLDNNs for voice activity detection,” in *Proc. of Interspeech*, 2016.

[12] H. Dinkel, N. Chen, Y. Qian, and K. Yu, “End-to-end spoofing detection with raw waveform CLDNNS,” in *Proc. of ICASSP*, 2017.

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

[14] S.-W. Fu, Y. Tsao, X. Lu, and H. 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] 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.

[16] D. Palaz, M. Magimai.-Doss, and R. Collobert, “Analysis of CNN-based speech recognition system using raw speech as input,” in *Proc. of Interspeech*, 2015.

[17] ——, “End-to-end acoustic modeling using convolutional neural networks for hmm-based automatic speech recognition,” *Speech Communication*, vol. 108, pp. 15–32, Apr. 2019.

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

[19] S. H. Kabil, H. Muckenhirn, and M. Magimai.-Doss, “On learning to identify genders from raw speech signal using CNNs,” in *Proceedings of Interspeech*, 2018.

[20] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” in *Proc. of International Conference on Learning Representations (ICLR)*, 2014.

[21] J. 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.

[22] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Proc. of European Conference on Computer Vision*, 2014.

[23] 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.

[24] T. Pellegrini and S. Mouysset, “Inferring phonemic classes from CNN activation maps using clustering techniques,” in *Proc. Interspeech*, 2016.

[25] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, “Visualizing higher-layer features of a deep network,” University of Montreal, Tech. Rep. 1341, Jun. 2009.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 2009, pp. 248–255.

[27] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *International Conference on Learning Representations*, 2015.

[28] W. Wirtinger, “Zur formalen theorie der funktionen von mehr komplexen veränderlichen,” *Mathematische Annalen*, vol. 97, no. 1, pp. 357–375, 1927.

[29] H. Caracalla and A. Roebel, “Gradient conversion between time and frequency domains using wirtinger calculus,” in *Proc. of International Conference on Digital Audio Effects*, 2017.

[30] P. Boersma, “Praat, a system for doing phonetics by computer.” *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[31] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. 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.

[32] T. Kinnunen, “Spectral features for automatic text-independent speaker recognition,” *Licentiates Thesis*, 2003.

[33] L. F. Gallardo, M. Wagner, and S. Möller, “Spectral sub-band analysis of speaker verification employing narrowband and wideband speech.”

[34] Ö. D. Orman and L. M. Arslan, “Frequency analysis of speaker identification,” in *2001: A Speaker Odyssey-The Speaker Recognition Workshop*, 2001.