Convolutional Neural Network with Semantically Meaningful Activations for Speech Analysis

Ryhor Vashkevich, Elias Azarov
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
ryhorv@gmail.com, azarov@bsuir.by

***Abstract*—Semantic analysis of speech is more prospective compared to analysis of text since speech contains more information that is important for understanding. The most important distinguishing feature of speech is intonation, which is inaccessible in the text analysis. For successful semantic analysis of speech it is necessary to transform the speech signal into features with semantic interpretation. The mathematical apparatus of convolutional neural networks (CNN) seems suitable to implement this kind of transformation. However there is a scalability problem that makes it hard to combine many CNN's in a single solution. To overcome this we propose to develop a CNN model with semantically meaningful activations i.e. the model that is capable of semantic interpretation of its internal states. The ultimate goal of the transform is to extract all semantically meaningful information, however the present work is confined to voice activity detection (VAD) and intonation extraction. Unlike other VADs based on artificial neural networks, the proposed model does not require a lot of computing resources and has a comparable or even better performance.**

***Keywords*—semantic speech analysis, voice activity detection, convolution neural network, VAD, CNN.**

I. INTRODUCTION

Known speech analysis and processing solutions based on neural networks can hardly be embedded into semantic systems because their internal states cannot be interpreted in semantic terms. In this paper we propose a CNN model that extracts speech intonation and voice activity using semantically meaningful activations.

A voice activity detector per se is one of the most important modules in many speech processing applications, such as audio coding, speech recognition, speaker identification, etc. The problem of voice detection in an audio signal has not yet been solved, especially in the presence of noises, which often present in the audio signal in the real world.

Significant development of machine learning in other speech processing tasks led to attempts to apply machine learning methods to VAD. In [1] the authors used a deep belief network (DBN) as the main tool for building their own VAD system. In [2, 3, 4] the authors used restricted Boltzmann machines (RBM) and networks with fully connected layers. A support vector machine (SVM) is used in [5, 6, 7] to classify features of a speech signal as one of the most computationally simple methods of classification. In [8, 9], the authors used the fact that a sound signal is a time series, and they used recurrent neural networks (RNN) for building VAD systems.

Another disadvantage of deep neural networks is computational complexity. Considering that the voice detector is often only an auxiliary module of a speech processing system, it is necessary to be sure that the VAD module consumes as little computing resources as possible.

The model proposed in this work provides a high accuracy of VAD comparable to existing solutions based on neural networks but uses much fewer (by several orders) parameters. A useful property of the obtained solution is the possibility of estimating a basic pitch of a speech signal. This estimation is generated by network activations.

II. PROPOSED METHOD

A. Features extraction

The choice of characteristic features of a sound signal is one of the most important part of a VAD system building process. We propose to use a fact that a speech signal has harmonic components, which our model tries to detect. As basic features of a speech signal, most works use mel-frequency cepstral coefficients (MFCC) [1, 4, 5, 6, 7, 8, 10]. Instead, we propose to use a spectrogram of an audio signal, and show that our model is able to efficiently detect the harmonic components of the signal, which are the main criterion for the presence of a voice in an audio signal. As shown in Fig. 1, the harmonic components of the sound signal are clearly visible in those parts of the spectrogram where the voice is present. In this case, the speech signal can be described using a fundamental frequency ($F0$), an amplitude, and number of harmonic components.

The fundamental frequency of the harmonic signal is the frequency corresponding to the first harmonic. The frequencies of all the harmonics of the speech signal are multiples of $F0$. To determine whether the signal is harmonic, we select the amplitudes of only those frequency components from the spectrogram that correspond to harmonics for a given $F0$ and feed them to the CNN model. In our experiments, we assume that $F0$ takes values from the range from $F0_{min} = 70$Hz and to $F0_{max} = 350Hz$. Also we introduce the notation for number of harmonics M and number of possible fundamental frequencies N. These variables are hyperparameters and may have different values.

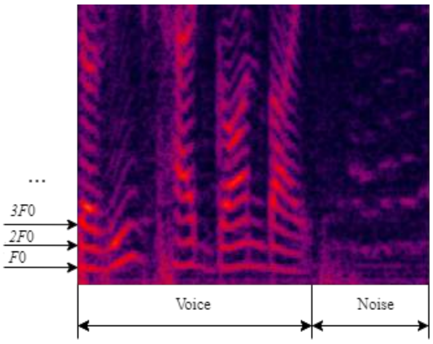

Figure 1. Spectrogram of sound signal.

The possible F0 range $[F0_{min}, F0_{max}]$ is uniformly covered by frequency grid G $[F0_0, \ldots, F0_{N-1}]$ where each point can be calculated as:

$$F0_i = Fo_{min} + i * \frac{F0_{max} - F0_{min}}{N-1}, i = \overline{0,(N-1)}$$

The indices of all the M-harmonics for given $F0_i$ are calculated as:

$$index_j = round\left(\frac{(1+j) * F0_i * N_{fft}}{2 * f_s}\right), j = \overline{0,(M-1)}$$

where f_s is audio sample rate, N_{fft} is the format of the fast Fourier transform (FFT).

Decision is frame-based (one frame one decision). The log amplitude spectrum of signal frame s is calculated as:

$$S = log_{10}|FFT(s)|$$

which is transformed into features vector X:

$$X(i,j) = S(index_j)$$

Thus for each signal frame we form a matrix of features X with shape $N \times M$, consisting of N points for M components. The basic idea is shown in Fig. 2.

We can consider each point of the frequency grid G as a pitch candidate. The task of the neural network model is to determine whether among the selected candidates there is one, which clearly represents the harmonic structure of the signal. If the model can detect a candidate that describes a harmonic signal, then the current input example corresponds to a speech signal, otherwise this example is classified as noise.

B. CNN architecture

In this paper we propose to use a simple model of a convolutional neural network consisting of only two 2D convolutional layers, followed by a global max-pooling layer. Fig. 3 shows the architecture of the proposed model.

The input of the model is a matrix of features X with shape $N \times M$. The first convolutional layer has K filters of size $1 \times M$ with a ReLu activation function. The second convolutional layer has only 1 filter with size 1×1. It aggregates the features selected by the convolution filters on the previous layer for each of the N candidates separately. This layer does not have an activation function. After than, the global max-pool layer selects the candidate with the maximum activation, thereby assuming that the candidate contains the harmonic component of the input signal. Further, the candidate selected passes through a sigmoidal activation function, as a result of which the output of the network will represent the probability that the input signal is a voice.

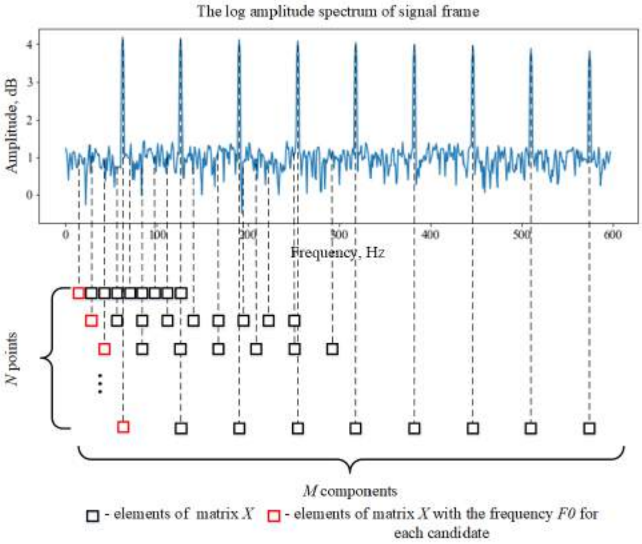
The log amplitude spectrum of signal frame
Amplitude, dB
4
3
2
1
0
0
100
200
300
400
500
600
Frequency, Hz
N points
M components
□ - elements of matrix X
□ - elements of matrix X with the frequency F0 for each candidate

Figure 2. The basic idea of features selection.

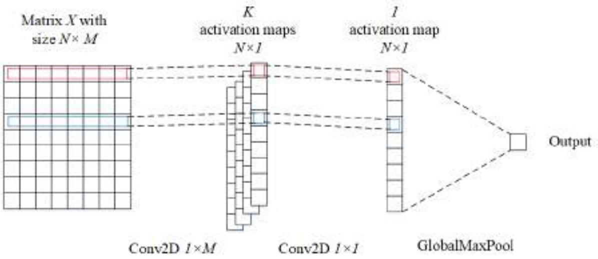

Figure 3. Architecture of CNN for VAD.

C. Pitch extraction

The second Conv2D layer forms an estimate of the pitch frequency. A high value of any activation of the second Conv2D layer gives a high degree of confidence that an input sample is a periodic signal.

III. Experiments

A. Training

The network was trained using SGD with learning rate η and momentum 0.9. We used a binary cross-entropy (1) as a loss function.

$$L_{ce} = \frac{1}{N_X} \sum_{i=0}^{N_x - 1} -y_i * log(t_i) - (1 - y_i) * log(1 - t_i) \quad (1)$$

where N_X is number of training samples, y_i is network output for i-th training sample, t_i is target value of the class label for the i-th training sample.

B. Dataset preparation

To train the model, we used our own dataset, consisting of 50101 examples. This dataset was divided into training and testing subsets in a ratio of 3 to 1. Additive and multiplicative components from white noise were added to the dataset.

C. Hyperparameters tuning

To determine optimal values for hyperparameters, we divided the dataset into training and testing subsets in a ratio of 3 to 1 and trained several model configurations with the different hyperparameter values. We experimented with different values of N, M and K. The hyperparameters tuning results are presented in Table I.

Table I
HYPERPARAMETERS TUNING RESULTS

Model	K	M	N	Training L_{ce}	Testing L_{ce}
M-3-7-100	3	7	100	0.61215	0.44838
M-3-7-50	3	7	50	0.59604	0.52501
M-3-7-200	3	7	200	0.63900	0.48520
M-3-14-100	3	14	100	0.63713	0.70615
M-3-14-50	3	14	50	0.59120	0.52208
M-3-14-200	3	14	200	0.56230	0.59663
M-3-20-100	3	20	100	0.69536	0.80170
M-3-20-50	3	20	50	0.57937	0.46226
M-3-20-200	3	20	200	0.61532	0.87776
M-10-7-100	**10**	**7**	**100**	**0.57937**	**0.37481**
M-10-7-50	10	7	50	0.60293	0.40301
M-10-7-200	10	7	200	0.89240	0.45877
M-10-14-100	10	14	100	0.59618	0.49430
M-10-14-50	10	14	50	0.65694	0.44893
M-10-14-200	10	14	200	0.92659	0.65541
M-10-20-100	10	20	100	0.62169	0.44377
M-10-20-50	10	20	50	0.61024	0.44842
M-10-20-200	10	20	200	0.88882	0.80149

As shown in Table I, the best results on the testing dataset has a model named "M-10-7-100". The optimal values of the all hyperparameters are presented in Table II. We used the hyperparameters of this model in the remaining experiments.

Table II
OPTIMAL HYPERPARAMETERS VALUES

Hyperparameter	Value
Number of candidates, N	100
Number of harmonics, M	7
Number of filters in the first Conv2D layer, K	10
Frame size, FFT size, N_{fft}	4096
Frame step	224
Signal sample rate, f_s	44100Hz
Learning rate, η	0.01

D. Comparison with other models

As a baseline model for comparison, we took the VAD model similar to that proposed in [11]. This model is a 4-layer neural network, consisting only of dense layers and uses MFCC, delta-MFCC, and delta-delta-MFCC as basic features. The input vector consists of $13 * 3 = 351$ nodes. Here, 13 is the number of MFCC coefficients, and 3 is the total number of different features.

The network was trained using SGD with learning rate 0.01, momentum 0.9, and loss function (1).

The training results of the proposed and basic models are presented in Table III. Both models were tested using the cross-validation technique.

Table III
RESULTS OF EXPERIMENTS

Criterion	Proposed model	Baseline model
Number of parameters	90	969,218
Training L_{ce}	0.33469	0.69316
Training accuracy, %	86.27	50.486
Testing L_{ce}	0.45104	0.69159
Testing accuracy, %	78.99	57.668

As show in Table III, our model has extremely less trainable parameters and it takes a lot less time to perform a forward pass.

E. Pitch extraction

As shown in Fig 4 our model can estimate the pitch frequency. The pitch frequency corresponds to the activation number of the second Conv2D layer. The model generates a hight output value if it can detect the pitch frequency of input sample.

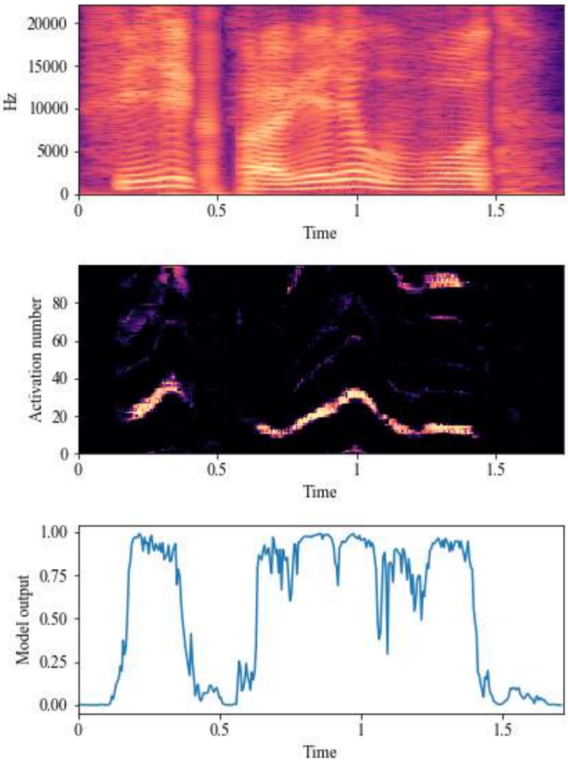

Figure 4. The pitch frequency estimation.

IV. Conclusion and future work

In this paper, we proposed a method for voice activity detection in a sound signal based on a simple convolutional neural network. A key feature of the method is the selection of characteristic features of a speech signal. Using the fact that a voiced speech has a harmonic structure as characteristic features we proposed to use sound signal spectrogram coefficients which are multiples of the specified fundamental frequency. Due to the fact that each voice has its own pitch frequency - we use 100 variants of values for the fundamental frequency. These values are uniformly located in the range from $70Hz$ to $350Hz$. The obtained features thus feed to the input of a 2-layer convolutional neural network, which classifies the input example into two classes - a voice or a noise.

Performance of the proposed model is comparable to state-of-the-art models based on neural networks, but our model contains significantly fewer trainable parameters. Therefore much less data is needed to train the model, and much less time is taken to perform a forward pass, and it increases the performance of the entire system.

Further improvement of the method will be aimed at 1) extending internal states of CNN to represent additional semantically important information; 2) increasing a model's inference quality using examples containing a harmonic signal which is not a voice (for example, musical instruments).

References

[1] Xiao-Lei Zhang, Ji Wu, Deep Belief Networks Based Voice Activity Detection, IEEE Transactions on Audio, Speech, and Language Processing, 2013, vol. 21, no. 4, pp. 697-710.

[2] Qing Wang1, Jun Du1, Xiao Bao, A universal VAD based on jointly trained deep neural networks, INTERSPEECH, 2015.

[3] T. G. Kang, K. H. Lee, W. H. Kang, DNN-based voice activity detection with local feature shift technique, Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016, pp. 1-4.

[4] Ryant, N., Liberman, M., Yuan, J., Speech activity detection on youtube using deep neural networks, INTERSPEECH, 2013, pp. 728-731.

[5] Tomi Kinnunen, Evgenia Chernenko, Marko Tuononen, Voice Activity Detection Using MFCC Features and Support Vector Machine, 2007.

[6] Y. X. Zou, W. Q. Zheng, W. Shi, Improved Voice Activity Detection based on support vector machine with high separable speech feature vectors, 19th International Conference on Digital Signal Processing, 2014, pp. 763-767.

[7] X. L. Zhang, J. Wu, Denoising deep neural networks based voice activity detection, IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 853-857.

[8] F. Vesperini, P. Vecchiotti, E. Principi, Deep neural networks for Multi-Room Voice Activity Detection: Advancements and comparative evaluation, International Joint Conference on Neural Networks (IJCNN), 2016, pp. 3391-3398.

[9] F. Eyben, F. Weninger, S. Squartini, Real-life voice activity detection with LSTM Recurrent Neural Networks and an application to Hollywood movies, IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 483-487.

[10] S. M. R. Nahar, A. Kai, Robust Voice Activity Detector by combining sequentially trained Deep Neural Networks, International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), 2016, pp. 1-5.

СВЁРТОЧНАЯ НЕЙРОННАЯ СЕТЬ С СЕМАНТИЧЕСКИ-ЗНАЧИМЫМИ АКТИВАЦИЯМИ ДЛЯ АНАЛИЗА РЕЧИ

Г. Вашкевич, И. Азаров
БГУИР
Минск, Беларусь

Семантический анализ речи является более перспективным по сравнению с анализом текста, поскольку речь содержит больше информации, которая важна для понимания. Самой важной отличительным признаком речи, недоступным текстовом анализе, является интонация. Для успешного семантического анализа речи необходимо из речевого сигнала выделить характеристические признаки с семантической интерпретацией. Математический аппарат свёрточных нейронных сетей (CNN) представляется подходящим для реализации такого рода преобразований. Однако существует проблема масштабируемости, которая затрудняет объединение нескольких CNN в одном решении. Чтобы преодолеть это, мы предлагаем разработать модель CNN с семантически значимыми активациями, то есть модель, внутрениие состояния которой можно интерпретировать с семантической точки зрения. Конечная цель преобразования состоит в том, чтобы извлечь из речи всю семантически значимую информацию, однако настоящая работа ограничивается детектированием голосовой активности и выделением интонации.

Благодатя предложенному в работе методу выделения характеристических признаков звукового сигнала и выбранной архитектуре нейронной сети стало возможным оценить частоту основного тона гармонического сигнала. Сильная активация какого-либо выхода второго слоя нейронной сети позволяет судить о гармонической природе входного сигнала. Если при этом сопоставить данный выход со шкалой частот, то можно будет получить численное значение частоты основого тона гармонического сигнала.

Предложенная модель по производительности сопоставима с другими современными моделями на основе нейронных сетей, однако содержит значительно меньше обучаемых параметров. Из этого следует, что для ее обучения необходимо гораздо меньше данных. При этом простота архитектуры нейронной сети позволяет использовать ее в мобильных платформах или встраиваемых системах.

Дальнейшее совершенствование метода будет направлено на повышение качества работы модели на примерах, содержащих гармонический сигнал, но при этом не относящийся к голосу (например звук музыкальных инструментов).