On Learning Vocal Tract System Related Speaker Discriminative Information from Raw Signal Using CNNs

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

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

hannah.muckenhirn@idiap.ch, mathew@idiap.ch,sebastien.marcel@idiap.ch

Abstract

In a recent work, we have shown that speaker verification systems can be built where both features and classifiers are directly learned from the raw speech signal with convolutional neural networks (CNNs). In this framework, the training phase also decides the block processing through cross validation. It was found that the first convolution layer, which processes about 20 ms speech, learns to model fundamental frequency information. In the present paper, inspired from speech recognition studies, we build further on that framework to design a CNN-based system, which models sub-segmental speech (about 2ms speech) in the first convolution layer, with an hypothesis that such a system should learn vocal tract system related speaker discriminative information. Through experimental studies on Voxforge corpus and analysis on American vowel dataset, we show that the proposed system (a) indeed focuses on formant regions, (b) yields competitive speaker verification system and (c) is complementary to the CNN-based system that models fundamental frequency information.

Index Terms: Speaker verification, convolutional neural network, end-to-end learning, fundamental frequency, formants

1. Introduction

For many years, state-of-the-art speaker recognition systems have been based on the extraction of handcrafted features relying on speech production and perception knowledge, such as Mel-frequency cepstral coefficients (MFCCs). During the training phase, such features are extracted from a large amount of speakers. A Gaussian Mixture model (GMM) is then trained to build a Universal Background Model (UBM). During the enrollment of a speaker, the mean of the GMM is adapted to fit the speaker's data. In GMM-UBM [1] system, the stacked mean vectors were directly used as the representation of the speaker. However, it has been shown that it is beneficial to process it further by extracting an i-vector from it [2]. During the verification phase, an i-vector is extracted from the given speech sample in the same manner. Then, it is compared to the reference i-vector, either with a simple cosine distance or with more complex techniques such as Probabilistic Linear Discriminant Analysis (PLDA) [3].

In recent years, a common trend has appeared in many pattern recognition fields. The principle is to remove as much as possible handcrafted processes and feed the raw signal to a deep neural network, which outputs directly the desired predictions. The motivation is that the neural network should be able to learn how to extract the relevant information from the signal in order to classify it, given that it is deep enough and that there is a sufficient amount of data. This approach has led to significant performance improvements in many fields such as image recognition [4] and natural language processing [5].

In speaker recognition, a similar trend has been observed. However, most "deep" systems are fed with intermediate features such as filter banks outputs [6, 7] and spectrograms [8, 9]. In other speech-related domains, such as speech recognition [10, 11, 12], emotion recognition [13], voice activity detection [14] and anti-spoofing [15, 16], it has been shown that neural networks fed with raw speech signal yield competitive systems. Inspired by these results, we recently proposed a speaker verification system where a Convolutional Neural Network (CNN) is trained directly on the raw speech signal [17]. The hyperparameters including block processing of speech signal is determined via cross validation. We found that this approach yields promising results and outperforms state-of-the-art systems on a subset of the Voxforge database, composed of 300 speakers. Furthermore, an analysis of first convolution layer, which processes the input signal with a kernel width of 300 samples ($\approx$ 20 ms), showed that the filters focus on low frequencies (below 1000 Hz) and model voice source related fundamental frequency information.

Speaker discriminative information, however, also lies at the vocal tract system level. This has been the underlying motivation for using cepstral features that model envelop of short-term spectrum for speaker recognition [18]. In speech recognition studies [10, 19], it has been found that CNNs with short kernel width of about 2 ms in the first convolution layer learns information related to formants, which is a vocal tract system related information. Inspired from these observations, in the present paper we investigate:

1. whether the use of short kernel width, similar to speech recognition studies, can model speaker discriminative information lying at the vocal tract system level?
2. how such a system compares to the CNN-based system that models voice source related information?

The paper is organized as follows. Section 2 describes the proposed CNN-based approach. Section 3 presents the experimental setup and the corresponding results obtained with the two proposed systems as well as with the baseline systems. Section 4 presents an analysis of the first layer of the CNNs.

2. CNN-based approach

Figure 1 illustrates the proposed approach with an architecture motivated from [10, 19] and developed for speaker recognition in our recent work [17]. The architecture consists of convolution layers followed by a multilayer perceptron (MLP). In this approach, the development of the speaker verification system consists of two steps:

1. in the first step, a CNN-based speaker identification with raw speech signal as input is trained to classify unknown speakers in an end-to-end manner. By unknown, we mean the speakers that are not part of the speaker ver-

ification system. This step is akin to UBM step in standard speaker verification approaches, except that here a speaker discriminative model as opposed to a generative model is trained.

2. in the second step, for each speaker s_m, $m = 1, \ldots, M$ in the speaker verification system, the CNN-based speaker identification system is converted into a speaker detection system for speaker s_m by: (a) replacing the output layer by two classes (genuine, impostor) and randomly initializing the weights between the output layer and the MLP hidden layer; and (b) adapting the CNN in a discriminative manner with the target speaker enrollment data and impostor speech data from unknown speakers.

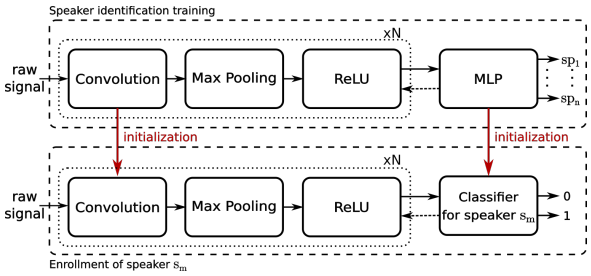


Figure 1: *Illustration of development of speaker verification system based on the proposed approach.*

In the verification phase, the test speech is passed through the speaker detection system of the claimed speaker and the decision is made by averaging the output posterior probability for genuine class and impostor class over time frames.
In our previous work, the architecture and the hyperparameters were determined via cross validation [17]. Through an analysis of the first convolution layer, it was shown that the CNN models fundamental frequency information. In the present paper, the focus is on designing the first convolution layer so as to model vocal tract system related speaker discriminative information and study such a system in terms of performance as well as information learned. In the following section, we first present a speaker verification study in that direction.

3. Speaker verification experiments

This section describes the experiments and the results obtained with our approach and with different baseline systems. All the experiments are reproducible.[1]

3.1. Database and experiment protocol

As done in our previous work [17], we perform speaker verification studies on Voxforge database. It is an open source speech database,[2] where different speakers have voluntarily contributed speech data for development of open resource speech recognition systems. The data and experiment protocol remains the same. More precisely, we selected 300 speakers who have recorded at least 20 utterances. We split this data into three subsets, each containing 100 speakers: the training, the development and the evaluation set, as described in Table 1. The 100 speakers with the largest number of recorded utterances are in the training set, while the remaining 200 were randomly split between the development and evaluation sets.

[1] https://gitlab.idiap.ch/biometric/CNN-speaker-verification-interspeech-2018

[2] http://www.voxforge.org/

Table 1: *Number of speakers and utterances for each set of the Voxforge database: training, development, evaluation.*

	training	development		evaluation	
		enrollment	probe	enrollment	probe
number of utterances/speaker	60-298	10	10-50	10	10-50
number of speakers	100	100		100	

The training set is used by the baseline systems to obtain a UBM. Whilst, it is used to obtain a speaker identification system in the proposed approach. The development and evaluation sets are split into enrollment and probe data. The enrollment data is used to train each speaker's model and contains 10 utterances per speaker. The probe part of the development data is used to fix the score threshold to achieve an Equal Error Rate (EER), while the Half Total Error Rate (HTER) is computed on the probe data of the evaluation set based on this threshold.

3.2. Systems

3.2.1. Baseline systems

We train several state-of-the-art systems on the Voxforge database using the spear toolbox [20]. We first perform a Voice Activity Detection (VAD), where frame-level energy values are computed, normalized and then classified into two classes. 60 dimensional MFCCs are then extracted from frames of 25ms shifted by 10ms (19 first coefficients with the energy + first derivative + second derivative). These features are then used as input to several state of the art systems: UBM-GMM [1], i-vectors [2] classified with a cosine distance or PLDA, inter-session variability (ISV) [21] and joint factor analysis (JFA) [22]. For all the aforementioned systems, we use the default parameters, previously tuned on a different subset of the Voxforge database, as presented in [20].

3.2.2. Proposed systems

The raw speech signal is split into N frames $x^1, \ldots, x^N$ of 10ms. The same VAD algorithm as for the baseline systems is applied. If the frame is classified as silent we discard it. Otherwise, we add to it c frames of context and normalize the resulting sequence to have zero mean and unit variance. This sequence of length of $(2c + 1) \times 10$ms is then fed to a CNN.

In the first step, the CNN-based speaker identification system was trained on the training data by splitting it into a training part (90%) and a validation part (10%) for early stopping. The proposed system has several hyper-parameters. These hyper-parameters were determined through a coarse grid search and based on validation accuracy. The best validation accuracy was obtained for an architecture with two convolution layers and one hidden layer in the MLP and a context $c = 25$ frames, i.e., the length of the input sequence is 510 ms. Each convolution layer is composed of 80 filters followed by a max pooling over 5 frames. In the first convolution layer, the kernel width $kW_1 = 300$ samples and the kernel shift, also called stride, $dW_1 = 10$ samples. In the second convolution layer, $kW_2 = 10$ and $dW_2 = 1$. The fully connected layer contains 100 hidden units. We refer to this system as "CNN $kW_1 = 300$".

As explained in Section 1, the main goal of this paper is to model vocal tract system information in the CNN-based speaker verification approach. For that, we take inspiration from CNN-based speech recognition studies [19], where it has been found that with sub-segmental speech signal (about 2 ms) as input the CNN is able to learn formant information. Thus, we train in

parallel a CNN with exactly the same architecture and hyper-parameters except that $kW_1 = 30$ samples instead of 300. We refer to this system as "CNN $kW_1 = 30$".

In the second step, we first developed speaker verification systems for the development set speakers and determined the threshold that yields the EER. We then developed speaker verification systems for the evaluation set speakers and evaluated them with the threshold found on the development set. In both cases, the enrollment data of each speaker was split into a training part (80%) and a validation part (20%) for adapting the CNN and MLP parameters discriminatively. The impostor examples were the same for all speakers in the development and evaluation sets and were obtained by randomly selecting 300 utterances from the training set, which was used to build the speaker identification system.

In all the cases, stochastic gradient descent based training with early stopping was performed with a cost function based on cross entropy using Torch software [23].

3.3. Results

Table 2 presents the HTER obtained with the baseline systems and the proposed CNN-based systems on the evaluation set of the Voxforge database. The system "CNN $kW_1 = 300$" yields a slightly improved performance than in [17] for two reasons: (i) the scheme to normalize the input speech signal is not the same. In the present work, we normalize the sequence fed to the CNN instead of the whole utterance and (ii) there are 80 filters in both convolution layers instead of 20.

We observe that "CNN $kW_1 = 300$" yields a slightly lower HTER than "CNN $kW_1 = 30$". However, both proposed CNN-based systems outperform the baseline systems. One possible reason is that the amount of enrollment data, which is on average ≈ 50 seconds per speaker, might not be sufficient for the baseline systems. A score level fusion of the two CNN-based systems, computed by simply taking the average, yields the best performance.

Table 2: *Performance of the baseline systems and the proposed CNN-based systems on the evaluation set.*

System	HTER (%)
UBM-GMM	3.05
ISV	2.40
i-vector, cosine distance	2.82
i-vector, PLDA	5.87
JFA	5.00
CNN $kW_1 = 300$	0.80
CNN $kW_1 = 30$	1.15
Fusion	**0.75**

4. Analysis: impact of first convolution layer kernel width

In this section, we analyze the spectral information that is being modeled by "CNN $kW_1 = 30$" and contrast it with the information that is being modeled by "CNN $kW_1 = 300$".

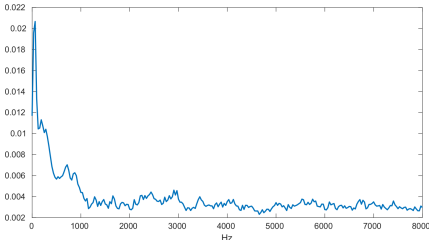
4.1. Visualization of filters

To understand the manner in which different parts of the spectrum are modeled, we analyze the cumulative frequency response of the learned filters similar to [19, 24]:

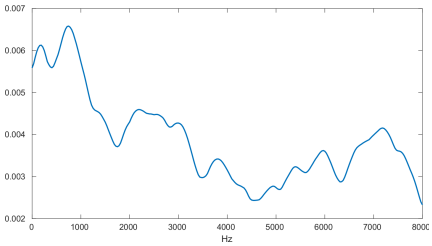
$$F_{cum} = \sum_{k=1}^{n_{f1}} \frac{\mathcal{F}_k}{\|\mathcal{F}_k\|_2}, \tag{1}$$

where n_{f1} is the number of filters in the first convolution layer and $\mathcal{F}_k$ is the magnitude spectrum of filter f_k, $k = 1, \ldots, n_{f1}$, computed with a 512-point Discrete Fourier Transform (DFT).

The resulting plots are shown in Figure 2. We observe that the cumulative responses are different. When $kW_1 = 300$ the filters give emphasis to low frequencies. On the other hand, when $kW_1 = 30$, the filters focus on different frequency regions. This indicates that the speaker discriminative information learned by the two systems are different.



(a) *System "CNN $kW_1 = 300$"*



(b) *System "CNN $kW_1 = 30$"*

Figure 2: *Cumulative frequency responses of first layer filters.*

4.2. Response of filters to input speech

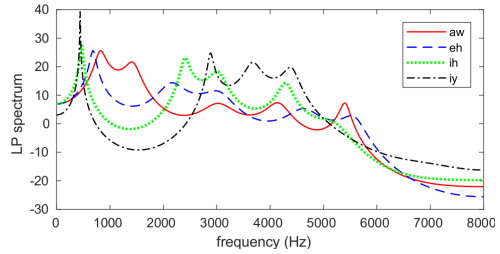
In the work on speech recognition [19], which formed the basis for the present work, it was found that the filters can be interpreted as a spectral dictionary, and the magnitude frequency response S_t of the input signal $s_t = \left\{s_t^1, \cdots s_t^{kW_1}\right\}$ can be estimated, as

$$S_t = \left| \sum_{k=1}^{n_{f1}} \langle s_t, f_k \rangle \text{DFT}\{f_k\} \right|, \tag{2}$$

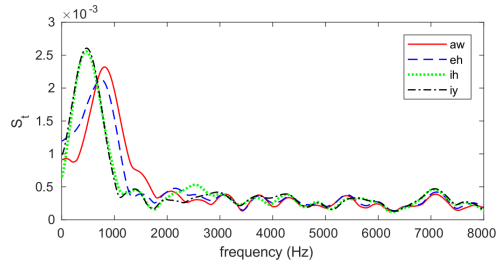
and analyzed to understand the discriminative information that is being modeled. If the filters f_k were to correspond to Fourier sines and cosines, then S_t would simply be the Fourier magnitude spectrum of s_t. In our previous work on "CNN $kW_1 = 300$", through such an analysis we showed that the first convolution layer learns to model fundamental frequency information [17].

Our hypothesis of using short kernel width, i.e., $kW_1 = 30$, is that in doing so vocal tract system information could be modeled for speaker discrimination. To ascertain that, we focus on analysis of formants, i.e., the resonances in the vocal tract system that change not only due to change in the speech sounds but also change due to speaker differences, such as different vocal tract lengths. So we performed an analysis on American English Vowel database [25]. This database contains recordings of 12 vowels uttered by 45 men, 48 women and 46 children with fixed context.

In Figure 3a, we show the linear prediction (LP) spectrum estimated for a frame of 30 ms speech signal for /aw/, /eh/, /ih/ and /iy/ produced by female speaker `w02`. The order of linear prediction is 20. In Figure 3b, we show the corresponding magnitude frequency response estimated based Eqn. (2) for exactly the same 30 ms frames. This is done by computing S_t after every 10 samples ($dW_1 = 10$ samples) in the 30 ms speech signal and averaging it.



(a) log *LP magnitude spectrum*



(b) *Magnitude frequency response* S_t

Figure 3: *Analysis of different vowels spoken by female speaker* `w02`.

In all cases we observe that the first spectral peak in the LP spectrum and the spectrum corresponding to the CNN-based system coincide. There is information related to the second formant in the spectral response of the CNN-based system but it is not clearly discernible. For example, for /ih/ the second peak could be seen as merger of second and third LP spectrum peaks. So we focus on the first spectral peak.

In order to ascertain that these observations generalize to other samples, we conducted a quantitative study on this database by comparing the first spectral peak locations with first formant location information provided with the database and available online [26] in the following manner:

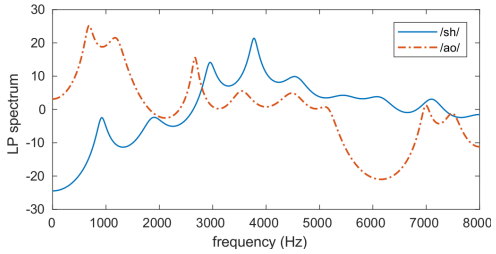
1. We extract the location (frequency) of the first peak of the LP magnitude spectrum.
2. We extract the first spectral peak location from the magnitude frequency response S_t of the CNN-based system.
3. We consider that the first formant location is correctly estimated if it is in the range $F_1 \pm (1 + \Delta)$, where F_1 is the value of the first formant.

We varied the Δ and computed accuracy over the whole database composed of 1668 utterances. Table 3 presents estimation accuracy for different values of Δ. We observe that for 83% of the samples the main peak of S_t is in the range $[0.85F_1, 1.15F_1]$ and is less precise than LP-spectrum. This indicates the CNN is focusing on speaker discriminative information present in the formant regions but is less precise about the speech sound.

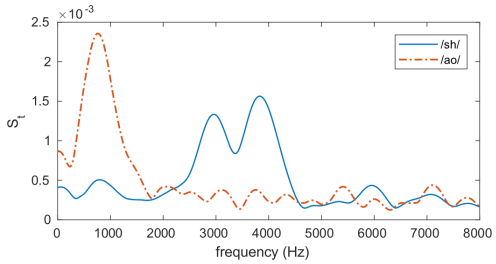
The analysis we conducted on the American English Vowel database gives a partial explanation of the filters response as we only consider vowels. In Figure 2, we can observe that the CNN is also modeling discriminative information above 2000 Hz. Figure 4 contrasts a frame of vowel /ao/ and a frame of fricative /sh/ produced by a speaker in TIMIT corpus when uttering word "wash". It can be observed that for /sh/ high frequency region is modeled.

Table 3: *Accuracy of first formant estimation in range* $[F_1(1 - \Delta), F_1(1 + \Delta)]$.

Δ	0.1	0.15	0.2
First peak of LP spectrum	0.93	0.97	0.98
First peak of CNN $kW_1 = 30$	0.55	0.83	0.93



(a) log *LP magnitude spectrum*



(b) *Magnitude frequency response* S_t

Figure 4: *Analysis of sound /oa/ and /sh/ extracted form word "wash" in TIMIT database.*

5. Conclusion

This paper focused on modeling vocal tract system related speaker discriminative information in a CNN-based end-to-end speaker verification system. Towards that, it investigated the use of short kernel width in the first convolution layer, more precisely, modeling sub-segmental (about 2ms) speech. Speaker verification studies showed that such a system yields performance comparable to the CNN-based system that models about 20 ms speech in the first convolution layer. An analysis of the trained CNNs showed that the two CNNs learn different information. Specifically, the CNN with short kernel width models formant regions for speaker discrimination as opposed to fundamental frequency information in the case of long kernel width. A system level combination indicates that the two approaches are complementary. Our future work will focus along two directions: (a) further investigations on challenging databases such as NIST SRE, VoxCeleb [9] and (b) development of i-vector based systems using the features learned by the CNNs.

6. Acknowledgment

This work was funded by the Swiss National Science Foundation through the project UniTS.

7. References

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, 2000.

[2] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[3] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. of International Conference on Computer Vision*, 2007.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012.

[5] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. of the International Conference on Machine Learning*, 2008.

[6] E. Variani, X. Lei, E. McDermott, I. Lopez 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.

[7] G. Heigold, I. Lopez Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proc. of ICASSP*, 2016.

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

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] ——, "Towards directly modeling raw speech signal for speaker verification using cnns," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, 2018.

[18] M. R. Sambur, "Selection of acoustic features for speaker identification," *IEEE Transactions on Audio, Speech, and Signal Processing*, vol. 23, no. 2, pp. 176–182, 1975.

[19] D. Palaz, M. Magimai.-Doss, and R. Collobert, "End-to-end acoustic modeling using convolutional neural networks for automatic speech recognition," Idiap, Idiap-RR Idiap-RR-18-2016, 6 2016, http://publications.idiap.ch/downloads/reports/2016/Palaz_Idiap-RR-18-2016.pdf.

[20] E. Khoury, L. El Shafey, and S. Marcel, "Spear: An open source toolbox for speaker recognition based on bob," in *Proc. of ICASSP*, 2014.

[21] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Computer Speech & Language*, vol. 22, no. 1, pp. 17–38, 2008.

[22] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[23] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A Matlab-like Environment for Machine Learning," in *BigLearn, NIPS Workshop*, 2011.

[24] 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.

[25] 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.

[26] [Online]. Available: https://homepages.wmich.edu/~hillenbr/voweldata.html