# TOWARDS DIRECTLY MODELING RAW SPEECH SIGNAL FOR SPEAKER VERIFICATION 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

## ABSTRACT

Speaker verification systems traditionally extract and model cepstral features or filter bank energies from the speech signal. In this paper, inspired by the success of neural network-based approaches to model directly raw speech signal for applications such as speech recognition, emotion recognition and anti-spoofing, we propose a speaker verification approach where speaker discriminative information is directly learned from the speech signal by: (a) first training a CNN-based speaker identification system that takes as input raw speech signal and learns to classify on speakers (unknown to the speaker verification system); and then (b) building a speaker detector for each speaker in the speaker verification system by replacing the output layer of the speaker identification system by two outputs (genuine, impostor), and adapting the system in a discriminative manner with enrollment speech of the speaker and impostor speech data. Our investigations on the Voxforge database shows that this approach can yield systems competitive to state-of-the-art systems. An analysis of the filters in the first convolution layer shows that the filters give emphasis to information in low frequency regions (below 1000 Hz) and implicitly learn to model fundamental frequency information in the speech signal for speaker discrimination.

***Index Terms***— Speaker verification, convolutional neural network, end-to-end learning, fundamental frequency

## 1. INTRODUCTION

The goal of a speaker verification system is to verify the identity claim of a person given an audio sample. Conventionally, short-term spectrum-based features such as Mel-Frequency Cesptral Coefficients (MFCCs) are modeled on a large set of speakers to build a universal background model (UBM) with a Gaussian mixture model (GMM). In UBM-GMM approach [1], the UBM is adapted on the speaker's data during the enrollment phase to obtain a speaker model. During the verification phase, the decision is made by a likelihood ratio test. The most popular approach is to extract a supervector by stacking the mean vectors of the speaker adapted GMM [2] and projecting the supervector onto a total variability space to extract a low dimensional representation called i-vector (identity vector) [3]. i-vector extraction is typically followed by a discriminative modeling technique, such as probabilistic linear discriminant analysis (PLDA) [4] to handle channel or session variation at the model level. In the verification phase, the decision is made through the PLDA score. The decision can also be made by simply computing a distance between i-vectors extracted during the enrollment phase and the verification phase.

This work was funded by the Swiss National Science Foundation through the project UniTS. The authors would like to thank Prof. B. Yegnanarayana for an informal discussion related to the analysis of filters.

In recent years, with the advances in deep learning, novel approaches are emerging where speaker verification systems are trained in an end-to-end manner [5, 6, 7]. These systems take as input either output of filterbanks [5, 6] or spectrograms [7, 8]. In this paper, we aim to go a step further where the features and the classifier(s) are learned by directly modeling the raw speech signal. Our motivation is two fold:

1. in recent works, it has been shown that raw speech signal can be directly modeled to yield competitive systems for speech recognition [9, 10, 11], emotion recognition [12], voice activity detection [13] and anti-spoofing [14, 15]. Can we achieve that for speaker recognition?
2. speaker differences occur at both voice source level and vocal tract system level [16, 17]. However, speaker recognition research has focused to a large extent on modeling features such as cepstral features and filter bank energies, which carry information mainly related to the vocal tract system, with considerable success. Can modeling of raw speech signal employing little or no prior knowledge provide alternate features or means for speaker discrimination?

Toward that, by building upon the end-to-end acoustic modeling approach for speech recognition presented in [9, 18], we develop a speaker verification approach where a convolution neural network (CNN) is first trained in an end-to-end manner to classify (unknown) speakers, and then adapted to build a speaker-specific binary classifier for speaker verification. Our investigations on the Voxforge corpus show that the proposed approach can yield systems competitive to state-of-the-art approaches. An analysis of the filters in the first convolution layer shows that the CNN gives emphasis to information lying in low frequencies (below 1000 Hz) and models fundamental frequency information.

The paper is organized as follows. Section 2 presents the proposed approach. Section 3 presents the experimental studies. Section 4 presents an analysis of what is learned by the first convolutional layer. Section 5 finally concludes.

## 2. PROPOSED APPROACH

Figure 1 illustrates the proposed approach with an architecture motivated from [9, 18], i.e., 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 verification 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.

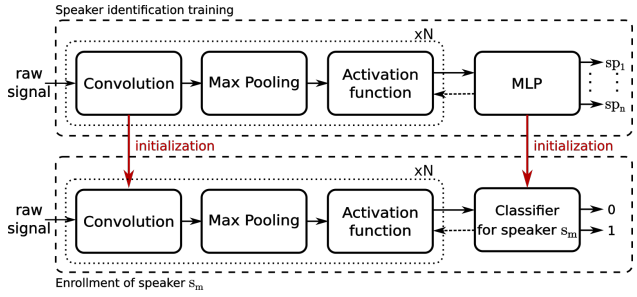



**Fig. 1**: Illustration of development of speaker verification system based on the proposed approach.

In the verification phase, the test speech is passed through the speaker detection system corresponding to the claimed speaker and the decision is made by averaging the output posterior probability for genuine class and impostor class over time frames.

The proposed system has the following hyper parameters: (i) window size of speech input ($w_{seq}$), (ii) number of convolution layers $N$, (iii) for each convolution layer $i \in \{1, \cdots N\}$, kernel width $kW_i$, kernel shift $dW_i$, number of filters $n_{fi}$ and max-pooling size $mp_i$ and (iv) number of hidden layers and hidden units in the MLP. All these hyper-parameters are determined through cross validation during the first step, i.e., development of the speaker identification system. In doing so, the system also automatically determines the short-term processing applied on the speech signal to learn speaker information. More precisely, the first convolution layer kernel width, i.e., $kW_1$ and kernel shift, i.e., $dW_1$ are the frame size and frame shift that operates on the signal. Figure 2 illustrates the first convolution layer processing. Note that the frame rate of the system is determined by the shift of input speech window of size $w_{seq}$, which was fixed at 10 ms, as done conventionally.

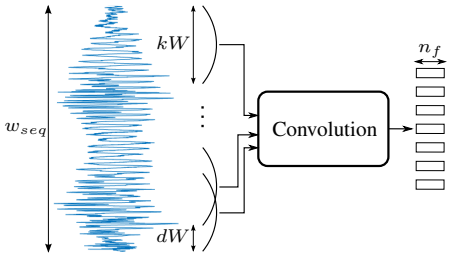



**Fig. 2**: Illustration of first convolution layer processing.

## 3. EXPERIMENTS

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

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

### 3.1. Database and experiment protocol

Voxforge is an open source speech database,[2] where different speakers have voluntarily contributed speech data for development of open resource speech recognition systems. Our main reason for choosing the Voxforge database was that most of the corpora for speaker verification have been designed from the perspective of addressing issues like channel variation, session variation and noise robustness. As a first step, our aim was to see whether the proposed approach could learn speaker discriminative information directly from the speech signal of short utterances, and if yes, whether we could analyze and find what kind of information. We can expect the Voxforge database to have low variability as the text is read and the data is likely to be collected in a clean environment as each individual records his own speech. However, the database consists of short utterances of about 5 seconds length recorded by speakers over the time.

From this database, 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. 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. The statistics for each set is presented in Table 1.

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

| | train | dev | | eval | |
|---|---|---|---|---|---|
| | | 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 data and probing data. The enrollment data is used to train each speaker's model and always contains 10 utterances per speaker. The probe part of the development data is used to fix the score threshold so as 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 [19]. 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 [3] classified with a cosine distance or PLDA, inter-session variability (ISV) [20] and joint factor analysis (JFA) [21]. For all the aforementioned systems, we use the default parameters, previously tuned on a different subset of the Voxforge database, as presented in [19].

#### *3.2.2. Proposed system*

The same VAD algorithm as for the baseline systems was employed to remove silent frames. Each utterance was then normalized by its mean and variance.

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

In the first step, the CNN-based speaker identification system was trained on the training data by splitting it into a train part (90%) and a validation part (10%) for early stopping. As discussed in Section 2, 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 of 4.52% at frame level and 0.31% at utterance level was obtained for an architecture with two convolution layers and one hidden layer in the MLP and following hyper-parameters: $w_{seq} = 510$ms, $n_{f1} = n_{f2} = 20$ filters, $kW_1 = 300$ samples, $dW_1 = 10$ samples, $kW_2 = 10$ frames/filter $\times\ n_{f1}$ filters $= 200$, $dW_2 = 1$ frame, $mp_1 = mp_2 = 5$ frames and $n_{hu} = 100$ hidden units. The frame rate at each convolution layer output is determined by $dW_1$ and $dW_2$.

In the second step, we first developed speaker verification systems for the development set speakers and determined the threshold that yields the EER. We obtained an EER of 1.18% for utterance-level average genuine class probability threshold of 0.267. 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 train 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 [22].

### 3.3. Results

Table 2 presents the HTER obtained with the baseline systems and the proposed CNN-based system on the evaluation set of the Voxforge database. We observe that the proposed system outperforms the baseline systems. One possible reason is that the amount of enrollment data, which is on average $\approx 50$ seconds per speaker, might not be sufficient for the baseline systems.

**Table 2**: Performance of the baseline systems and the proposed CNN-based system 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** | **1.20** |

## 4. ANALYSIS

This section presents an analysis to get insight about the speaker discriminative information that is getting modeled at the first convolution layer of the speaker identification system.

### 4.1. Visualization of filters

To understand the manner in which different parts of the spectrum are modeled, we analyzed the cumulative frequency response of the learned filters similar to [18, 23]:

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

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

The resulting plot is shown in Figure 3. We can observe that the filters are giving emphasis to the information lying below 1000 Hz.

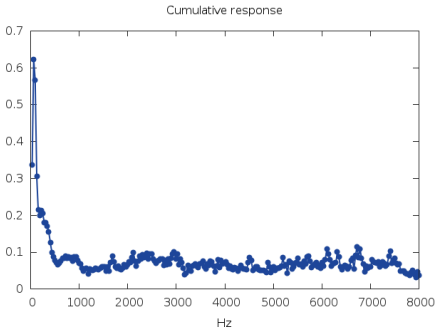



**Fig. 3**: Cumulative frequency response of filters of the first layer.

### 4.2. Response of filters to input speech

The previous analysis shows what frequency regions the filters are modeling but not how the filters respond to input speech. In the work on speech recognition [18], which formed the basis for the present work, it was found that the filters can be interpreted as a spectral dictionary,[3] 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

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

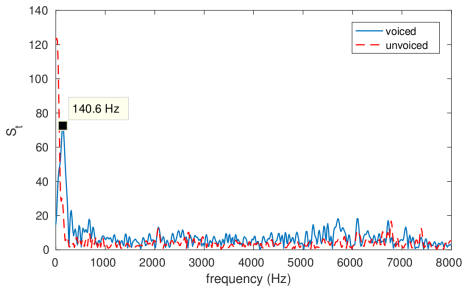
and analyzed to understand the discriminative information that is being modeled.

We adopted that approach to understand the speaker discriminative information that is getting modeled. In our case $kW_1 = 300$ speech samples and $n_{f1} = 20$. In the previous section, we observed that the filters are giving emphasis to low frequency information. One of the speaker-specific information that lies below 500 Hz is fundamental frequency. Considering this point we performed analysis of voiced speech and unvoiced speech of a few male and female speakers in the development set. In the case of voiced speech, we found a distinctive peak occurring at the fundamental frequency ($F_0$), while no such distinctive peak appears for unvoiced speech. Figure 4 illustrates that for two voiced and two unvoiced frames belonging to two different speakers. In both voiced speech cases, a distinctive peak is present in the frequency response near the corresponding $F_0$ values. Whilst, in the unvoiced speech cases the energy is very low in the region corresponding to $F_0$ range 70 Hz - 400 Hz compared to the voiced speech case. This suggests that the first convolution layer is learning $F_0$ modeling.

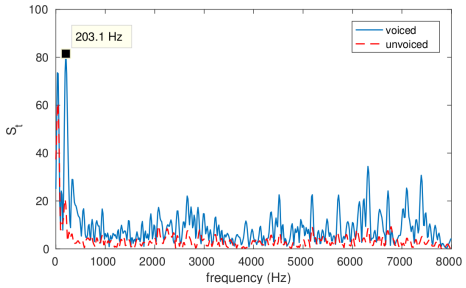
### 4.3. $F_0$ estimation using convolution filters

In order to ascertain that the first convolution layer is indeed learning to determine $F_0$, we implemented a simple $F_0$ estimator based on the observations made in the previous section and evaluated it on the Keele Pitch database [27], which contains the speech and laryngograph signal for 5 male and 5 female speakers reading a phonetically balanced text as well as hand corrected $F_0$ estimates from the

[3] It is worth mentioning that such interpretations have also been put forward in the signal processing community [24, 25].




(a) $F_0 = 149$ Hz for the voiced frame input, estimated using wavesurfer [26].




(b) $F_0 = 206$ Hz for the voiced frame input, estimated using wavesurfer.

**Fig. 4**: Filters response for voiced and unvoiced male speech and female speech frame inputs.

laryngography signal. The steps involved in the $F_0$ estimation are as follows:

1. For each frame of input signal of length $kW_1 = 300$ samples, estimate the frequency response $S_t$ using Eqn. (1) given the convolution filters parameters.
2. Locate the DFT bin that has maximum energy in the frequency range 70 Hz - 400 Hz.
3. Threshold the peak energy to decide if the frame is voiced or unvoiced. If voiced then the frequency corresponding to the DFT bin is the $F_0$ estimate.
4. Apply a median filter on the estimated $F_0$ contour.

The speech was down-sampled from 20 kHz to 16 kHz to match the sampling frequency of the Voxforge database. The frame shift was set to 10ms, as done in the Keele database for determining the reference $F_0$ from the laryngograph signal. The number of points for DFT was set as 3092 points. The energy threshold to decide voiced/unvoiced and the size of median filter were determined on the female speaker `f1n0000w` speech, such that low voiced/unvoiced (V/UV) error and gross error (i.e. deviation of estimated $F_0$ is within 20% of reference $F_0$ or not) is obtained. This threshold and the median filter size (=7) was used when estimating $F_0$ contours of the remaining nine speakers data and for evaluating the $F_0$ estimator. Figure 5 shows the $F_0$ contours for the first phrase spoken by a female and a male speaker. It can be observed that the estimated $F_0$ contours are reasonably close to the reference $F_0$ contours.

Table 3 presents the results of the evaluation. As it can be seen, the performance of this simple $F_0$ estimator is clearly beyond chance-level performance. The estimation for females are better than for males. The reason for this could be that frames of $kW_1 = 300$ samples, which amounts to 19ms, do not contain enough pitch cycles for very low $F_0$. We have indeed observed that through informal analysis of the errors.

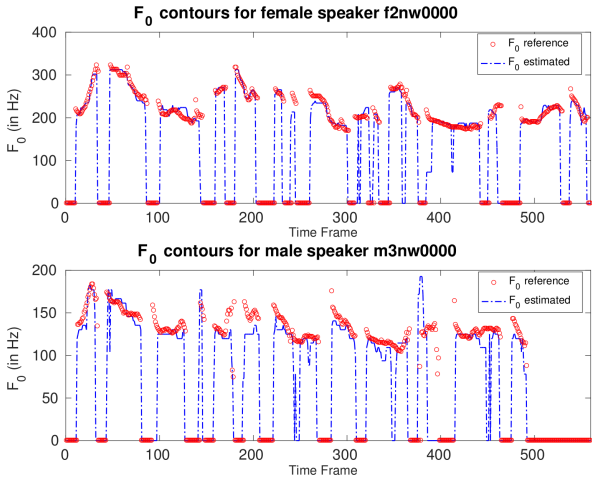



**Fig. 5**: Two examples of the $F_0$ contours estimated using the first layer filters compared to the reference $F_0$ from the database.

**Table 3**: $F_0$ estimation evaluation on the Keele database

| | V/UV error (%) | Gross error (%) |
|---|---|---|
| female (male) | 16.1 (22.3) | 3.6 (24.0) |

## 5. DISCUSSION AND CONCLUSION

In this paper, we proposed a speaker verification approach that learns speaker discriminative information directly from the raw speech signal using CNNs in an end-to-end manner. On the Voxforge corpus, the proposed approach yielded a system that outperforms systems based on state-of-the-art approaches. An analysis of the filters in the first convolution layer revealed that the filters give emphasis to information present in low frequency regions. Furthermore, an investigation on the response of the filters to input speech showed that the first convolution layer is implicitly learning to model $F_0$ for speaker discrimination. These observations together with the fact that the input to the system is 510 ms speech and the second convolution layer temporally filters and combines the first convolution layer filter outputs suggest that the system is learning to discriminate speakers based on "supra-segmental" information such as intonation patterns. These findings open interesting research questions:

1. we found that the system focuses on low frequencies and models $F_0$, which is a voice source related speaker discriminative information. A natural question that arises is: what other speaker discriminative voice source related information is it capturing? For instance, is it capturing voice quality related information?
2. in the present study the first convolution layer kernel width $kW_1$ determined in a cross validation manner is about 19 ms speech (segmental), and we found that it is modeling voice source related information. In the work on speech recognition [18], $kW_1$ determined in a cross validation manner was about 2 ms (sub-segmental), and was found to model formant-like information, which is related to vocal tract system. So can short $kW1$, i.e., modeling sub-segmental speech, help in capturing vocal tract system related speaker differences prominently and understanding them better?

Our future work will address these questions along with investigations on other corpora such as MOBIO [28] and NIST that have high variability in terms of channel and sessions.

## 6. REFERENCES


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

[2] William M Campbell, Douglas E Sturim, and Douglas A Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

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

[4] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. of International Conference on Computer Vision*, 2007.

[5] 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.

[6] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer, "End-to-end text-dependent speaker verification," in *Proc. of ICASSP*, 2016.

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

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] Jared J. Wolf, "Efficient acoustic parameters for speaker recognition," *The Journal of the Acoustical Society of America*, vol. 51, no. 6B, pp. 2044–2056, 1972.

[17] Marvin 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.

[18] 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, `http://publications.idiap.ch/downloads/reports/2016/Palaz_Idiap-RR-18-2016.pdf`.

[19] Elie Khoury, Laurent El Shafey, and Sébastien Marcel, "Spear: An open source toolbox for speaker recognition based on bob," in *Proc. of ICASSP*, 2014.

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

[21] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre 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.

[22] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet, "Torch7: A Matlab-like Environment for Machine Learning," in *BigLearn, NIPS Workshop*, 2011.

[23] 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.

[24] Vardan Papyan, Yaniv Romano, and Michael Elad, "Convolutional neural networks analyzed via convolutional sparse coding," *Journal of Machine Learning Research*, vol. 18, no. 83, pp. 1–52, 2017.

[25] Stéphane Mallat, "Understanding deep convolutional networks," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, pp. 20150203, 2016.

[26] Kåre Sjölander and Jonas Beskow, "Wavesurfer-an open source speech tool," in *Proc. of International Conference on Spoken Language Processing*, 2000.

[27] Fabrice Plante, Georg F Meyer, and William A Ainsworth, "A pitch extraction reference database," in *Proc. of EuroSpeech*, 1995.

[28] Chris McCool et al., "Bi-modal person recognition on a mobile phone: using mobile phone data," in *Proc. of Workshop on Hot Topics in Mobile Multimedia*, 2012.