On Learning to Identify Genders from Raw Speech Signal using CNNs

Selen Hande Kabil[1,2]*, Hannah Muckenhirn*[1,2]*, Mathew Magimai.-Doss*[1]

[1]Idiap Research Institute, Martigny, CH
[2]École Polytechnique Fédérale de Lausanne, CH

skabil@idiap.ch, hmuckenhirn@idiap.ch, mathew@idiap.ch

Abstract

Automatic Gender Recognition (AGR) is the task of identifying the gender of a speaker given a speech signal. Standard approaches extract features like fundamental frequency and cepstral features from the speech signal and train a binary classifier. Inspired from recent works in the area of automatic speech recognition (ASR), speaker recognition and presentation attack detection, we present a novel approach where relevant features and classifier are jointly learned from the raw speech signal in end-to-end manner. We propose a convolutional neural networks (CNN) based gender classifier that consists of: (1) convolution layers, which can be interpreted as a feature learning stage and (2) a multilayer perceptron (MLP), which can be interpreted as a classification stage. The system takes raw speech signal as input, and outputs gender posterior probabilities. Experimental studies conducted on two datasets, namely AVspoof and ASVspoof 2015, with different architectures show that with simple architectures the proposed approach yields better system than standard acoustic features based approach. Further analysis of the CNNs show that the CNNs learn formant and fundamental frequency information for gender identification.

Index Terms: automatic gender recognition, convolutional neural networks, multilayer perceptron, end-to-end training

1. Introduction

Automatic Gender Recognition (AGR) task focuses on identifying the speaker gender given the speech signal. AGR systems are useful for different speech applications, such as reducing search space in speaker recognition, building gender specific acoustic models for automatic speech recognition and understanding human-computer interactions.

Acoustic differences in the speech signal due to gender can be mainly attributed to two physiological aspects of the speech production system [1, 2]. More precisely, one related to the voice source. Males typically have lower fundamental frequency than females. This is mainly due to differences in the size of the vocal folds. The other related to the vocal tract system. Males typically have longer vocal tract than females. As a consequence formant frequency locations shift. Building on this point, typically in the literature [3, 4, 5, 6, 7, 8], two broad classes of features are used for this task: fundamental frequency (F0) and short term features like mel frequency cepstrum coefficients (MFCCs). There are also works that have investigated high level representations like Gaussian mixture model supervector [9, 8] and i-vectors [10].

As for the classifiers used to classify these features in AGR task, in literature, logistic regression, linear regression, random forests and support vector machines are employed [6, 7, 8, 9]. In [6], it is indicated that random forest trained on simple F0 and MFCC features performs close to the state-of-the-art system devised for 3-way classification problem (between male, female and child speech), which is a fusion of six subsystems.

In this paper, rather than approaching the problem of gender recognition in a divide and conquer manner, we aim to develop end-to-end automatic gender recognition system that learns both the relevant features and the classifier directly from the raw speech signal. This is motivated for recent successes in directly modeling raw speech signal for various tasks, such as speech recognition [11, 12, 13], emotion recognition [14], voice activity detection [15], presentation attack detection [16], speaker recognition [17]. In particular, we build upon recent works [12, 16, 17] to investigate the following:

1. how well such an end-to-end AGR system based on convolutional neural networks (CNNs) can identify genders, when compared to short-term spectral feature and fundamental frequency based AGR systems?
2. what kind of information the CNN models for AGR?

Towards that we present investigations on two corpora, namely, AVspoof and ASVspoof 2015. Through in-domain and cross-domain studies we demonstrate the potential of the approach.

The remainder of the paper is organized as follows. Section 2 presents the proposed approach. Section 3 and Section 4 presents the experimental setup and results, respectively. Section 5 presents an analysis of the trained CNNs. Finally, we conclude in Section 6.

2. Proposed Approach

Figure 1 illustrates the proposed approach. The proposed architecture is composed of several filter stages, followed by a classification stage. This architecture was first proposed for speech recognition [11, 12] and later successfully applied for presentation attack detection [16] and speaker verification [17].

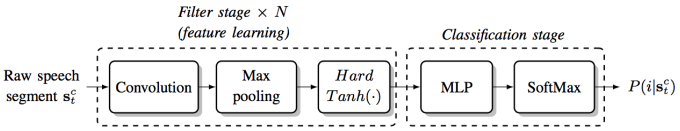

Figure 1: *Overview of convolutional neural network architecture.*

The input to the system s_t^c is raw speech signal, with the context of c frames ($s_t^c = s_{t-c} \ldots s_t \ldots s_{t+c}$) and the output of the system is the posterior probability $P(i \mid s_t^c)$, the probability of the gender class given the speech segment.

The filter stage has several hyper-parameters namely, kW, the temporal window width as an input to each convolutional layer, dW, the shift of the temporal window at each convolutional layer, mp, the max pooling kernel width, c, the context, n_f, the number of filters in convolutional layer and Lr, the learning rate. In addition, $clNhu$ is the hyper-parameter for the number of hidden units in classification stage.

The feature and classifier stages are jointly trained using stochastic gradient descent algorithm with cross entropy based error criterion. All the hyper-parameters are determined during the training, based on frame level classification accuracy on validation data.

3. Experimental Setup

This section first provides a description of the databases and protocols, and then describes the systems developed.

3.1. Databases and protocols

We conduct AGR studies on subsets of genuine samples extracted from two databases designed for spoofing tasks: the Audio-Visual Spoofing (AVspoof) [18] database and the Automatic Speaker Verification Spoofing (ASVspoof) 2015 [19] database.

The splitting of the subsets of AVspoof was the same as the one originally provided, while we were constrained to modify the protocol of the ASVspoof database, as the gender information was not given for the evaluation set. We thus used the original development set as our evaluation set and randomly split the speakers of the original training set to create a training and development set. The distribution of the number of speakers and utterances for datasets is shown in Table 1.

Table 1: *Distribution of the number of speakers and utterances for AVspoof and ASVspoof databases on training, development and evaluation sets.*

Database	set	speakers		utterances		style		
		female	male	female	male	pass	read	free
AVspoof	train	4	10	1504	3469	835	2900	1238
	dev	4	10	1434	3561	840	3020	1135
	eval	5	11	1722	3854	964	3412	1200
ASVspoof	train	20	15	1999	1498	-	-	-
	dev	7	5	700	500	-	-	-
	eval	26	20	5351	4053	-	-	-

The training set was used to optimize the parameters of the classifier. The evaluation was based on the performance of our system with Recognition Rate (RR) computed on the evaluation set. More precisely, frame level gender probabilities are combined over the utterance and final decision is made based on the utterance level probability score.

3.2. Systems

We first present a description of the baseline systems and then the proposed systems.

3.2.1. Baseline systems

For the baseline system, we adopted divide and conquer approach. That is, we first extracted the features from the speech signal, and then passed them to the classifier. We used the HTK toolkit [20] for cepstral feature extraction and snack toolkit [21] for fundamental frequency extraction. The MFCC features were computed with a frame size of 25 ms and a frame shift of 10 ms. The fundamental frequency features were computed with a frame size of 40ms and a frame shift of 10ms. We used quicknet tool [22] to train artificial neural networks (ANNs) with one hidden layer and various number of hidden units (50,100,250,500,750,1000). The ANN with best accuracy on the cross validation data was selected.

Two sets of baseline systems were developed. First, baseline system that models 38 dimensional MFCC features with four frames preceding and four frames following context (input dimension 38x9 = 342). We refer to this system as *ANN1*. Second baseline system that models 38 dimensional MFCC features and one dimensional fundamental frequency features with four frames preceding and four frames following context (input dimension 39x9 = 351). We refer to this system as *ANN2*.

3.2.2. Proposed systems

For the proposed system, we trained the CNN-based $P(i \mid s_t^c)$ estimator using raw speech signal. As we tried to recognize the gender from the given speech signal, s_t^c was the raw speech segment and i was the gender label (female, male). The only preprocessing step was to remove the silent frames at the beginning and at the end of the speech sample with an energy-based voice activity detection algorithm. Each sequence s_t^c fed to the CNN was normalized by removing the mean (of the sequence s_t^c) and dividing (each value in sequence s_t^c) by standard deviation. Torch7 toolbox [23] was used for training the systems.

Table 2: *The ranges of hyper-parameters for the grid search. Note that in our architectures, only one hidden layer MLPs with various units ($clNhu$) are used.*

Parameters	**Units**	**Range**
Context (c)	frames	5-20
Kernel width of first convolution ($kW1$)	samples	10-1500
Kernel shift of first convolution ($dW1$)	samples	10-500
Number of filters in first conv (n_f1)	filters	5-100
Max Pooling kernel width ($mp1$)	frames	2-10
Number of hidden units in classifier ($clNhu$)	units	10-200
Learning rate (Lr)		0.0001-0.01

As detailed in Section 2, the hyper-parameters that needed to be set were: c, n_f, dW, kW, mp, $clNhu$, Lr. These hyper-parameters were chosen based on the frame-level performance achieved on the development set during the training phase on AVspoof database. The hyper-parameter ranges which were considered as coarse grid search are shown in Table 2. The hyper-parameter values for selected three architectures are shown in Table 3.

Table 3: *The fixed hyper-parameters for the one convolution layered cnn1, two convolution layered cnn2 and three convolution layered cnn3 architectures*

arch	kW	dW	n_f	$clNhu$	c	Lr	mp
cnn1	360	20	20	20	15	0.0001	5
cnn2	360/65	20/1	20/10	20	15	0.0001	2/2
cnn3	60/25/12	10/5/2	100/10/2	50	15	0.0001	2/2/2

4. Results

We performed two set of studies: (i) in-domain study where the systems are trained and tested on the same corpus. We denote the in-domain studies as AVspoof and ASVspoof and (ii) cross-domain study where the system is trained on one corpus and tested on another corpus. In other words, training the AGR system on AVspoof data and testing on ASVspoof. Similarly training the AGR system on ASVspoof data and testing on AVspoof data. We denote these studies as AVtrain-ASVtest and ASVtrain-AVtest. Table 4 presents the results for in-domain and cross-domain studies.

In the in-domain studies, we can observe that all the three proposed systems outperform MFCC-based (*ANN1*) and MFCC+F0 based (*ANN2*) system. Comparison across the CNN-based systems shows that two convolution layers are needed to effectively identify gender. The performance with long kernel width (cnn2) and short kernel width (cnn3) are comparable.

Table 4: *Comparison of the baseline systems and the proposed systems in terms of RR*

Systems	AVspoof	ASVspoof	AVtrain-ASVtest	ASVtrain-AVtest
ANN1	86.0	93.5	55.8	82.8
ANN2	86.9	83.6	54.9	76.3
cnn1	91.3	99.6	89.7	39.0
cnn2	98.5	99.8	94.7	53.2
cnn3	97.2	99.8	89.4	39.2

In the cross-domain studies, the performance of the systems degrade. We observe two opposite trends. More precisely, for standard acoustic feature based system, the system trained on ASVspoof data generalizes better than the system trained on AVspoof data. Whilst for the proposed systems, the system trained on AVspoof generalizes better than the system trained on ASVspoof. The poor generalization of the baseline systems trained on AVspoof data could potentially be due to low number of speakers in the training set. The poor generalization of CNN based systems trained on ASVspoof data could be due to acoustic mismatch. ASVspoof data is collected in anechoic recording chamber while AVspoof data is collected in realistic conditions with multiple devices. An examination of the results showed that the performance was poor for the portion of AVspoof test data which was collected with smartphones. Together these results indicate that the proposed approach can yield robust system with less number of speakers but is sensitive to acoustic mismatch.

Overall, in both in-domain and cross-domain studies, *ANN1* yields better baseline. Whilst cnn2 consistently yields the best system across all the systems, except for ASVtrain-AVtest.

5. Analysis

In this section, we first analyze the frequency response of the filters learned in the first convolution layer and then analyze the response these filters to the input speech.

5.1. Cumulative Frequency Response of Learned filters

We analyze the spectral regions that are being modeled by the filters in the first convolution layer by computing cumulative frequency response, similar to [12]:

$$F_{cum} = \sum_{m=1}^{M} \frac{F_m}{\|F_m\|_2} \quad \text{, M: number of filters} \qquad (1)$$

where F_m is the magnitude spectrum of filter m.

Figure 2 presents the cumulative response for cnn2 and cnn3 trained on AVspoof corpus. It can be observed that for cnn2 the main emphasis is on low frequencies, while for cnn3 the emphasis is spread across frequencies, including low frequencies. These differences in the cumulative frequencies can be attributed to the different kernel widths used by these systems.

5.2. Response of the filters to input speech signal

In order to understand how these filters respond to input speech, we performed an analysis using speech from American English Vowels dataset [24]. American English Vowels dataset consists of recordings for 12 vowels (/ae/, /ah/, /aw/, /eh/, /er/, /ey/, /ih/, /iy/, /oa/, /oo/, /uh/, /uw/), for each of the speakers (50 men, 50 women, 29 boys, 21 girls). In addition, in [25], the frequency range for the F0 and formants were calculated and presented for each utterance in the dataset.

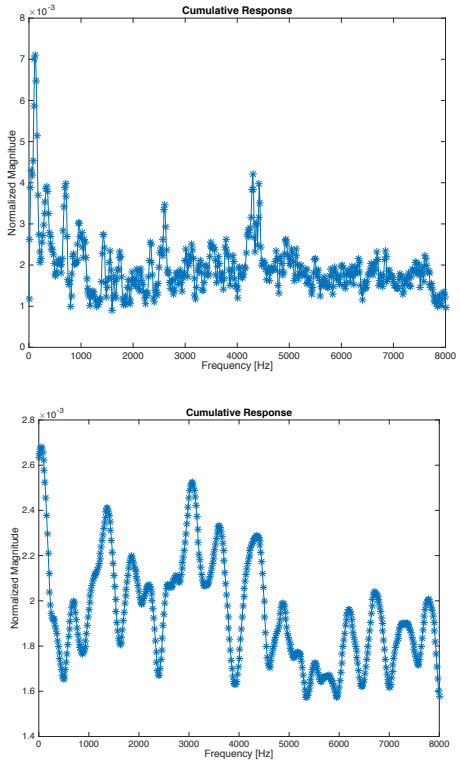

Figure 2: *Cumulative frequency responses of first convolution layer of cnn2 (top) and cnn3 (bottom) trained on AVspoof corpus.*

The response of the filters to the input speech was calculated in the following manner:

1. s_t^c was taken as the input speech segment. For the sake of simplicity, a window size of 30 ms similar to the one used in standard short term processing, was used.
2. The successive windows of kW samples (360 samples for cnn2 and 60 samples for cnn3) interspaced by dW samples (20 samples for cnn2 and 10 samples for cnn3) were taken from s_t^c.
3. For each of these successive window signals (s_t), the output of the filters y_t to the input speech signal s_t = $s_{t-(kW-1)/2}$ ... $s_{t+(kW-1)/2}$ was estimated as
$$y_t[m] = \sum_{l=-(kW-1)/2}^{l=+(kW-1)/2} f_m[l].s_{t+l} \qquad (2)$$
where f_m denotes the m^{th} filter in first convolution layer and $y_t[m]$ denotes the output of the m^{th} filter at time frame t.
4. The frequency response S_t of the input signal s_t was estimated as
$$S_t = |\sum_{m=1}^{M} y_t[m].\mathcal{F}_m|, \qquad (3)$$
where $\mathcal{F}_m$ is the complex Fourier transform of filter f_m.
5. average filter response for the input speech segment (s_t^c) was calculated by summing all frequency responses at all frames and by dividing it to the number of successive windows in the input speech segment.

The spectrum estimation process was originally developed in [12] and has been applied in other studies such as [17] to understand the frequency information captured by the CNNs. We performed filter response analysis for different vowels and speakers.

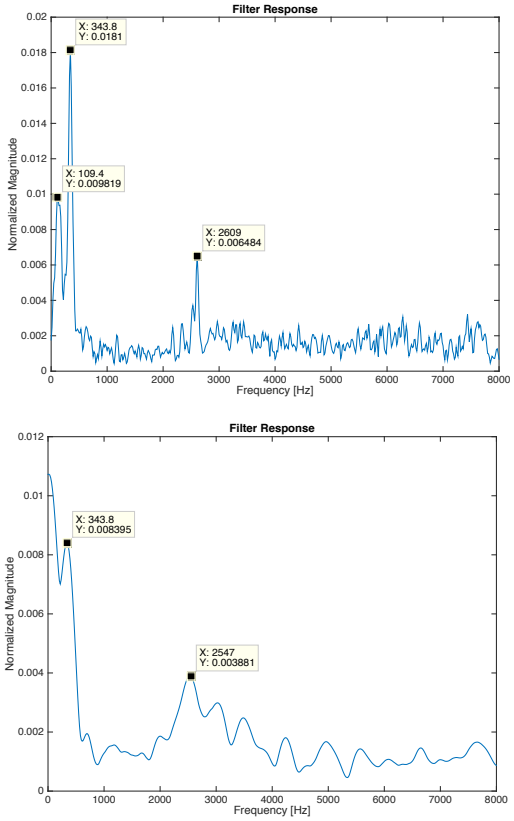

Figure 3: *Average filter response for 30ms speech of /iy/ utterred by male speaker* `m01` *for cnn2 (top) and cnn3 (bottom).*

Table 5: *The value range for F0,F1 and F2 (in Hz) for phone /iy/ utterances in American English Vowels dataset [25].*

Utterance	F0 interval	F1 interval	F2 interval
`m01iy`	96-216	305-402	2049-2600
`w10iy`	155-275	331-531	2129-2654

Figure 3 and Figure 4 shows the filter response of 30 ms speech of /iy/ uttered by male speaker `m01` and female `w10`, respectively. In the figures, information such as fundamental frequency F0, first formant F1 and second formant F2, which clearly appear as spectral peaks have been marked. For reference purpose, Table 5 presents the F0 interval, F1 interval and F2 interval that were established in [25] for the utterances `m01iy` and `w10iy`. It can be observed that cnn2, which has a kernel width of 360 samples ($\approx$ 22 ms), models both F0 and formant information for gender identification. Whilst cnn3, which has a kernel width of 60 samples ($\approx$ 4 ms), models only formant information. This explains why cnn2 yields a better system than cnn3. We have observed similar trends across different vowels and speakers. A quantitative analysis is in progress. It is worth mentioning that short kernel width CNN modeling only formant information is consistent with speech recognition studies [12], while long kernel width CNN modeling fundamental frequency

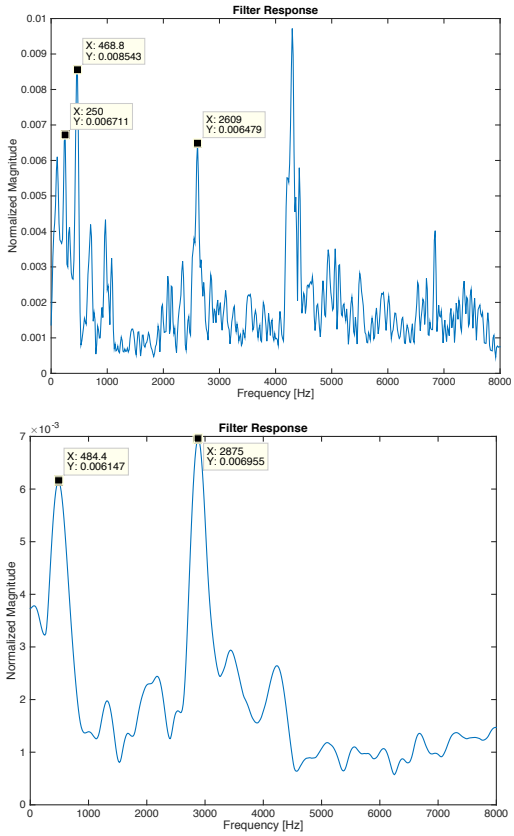

Figure 4: *Average filter response for 30ms speech of /iy/ utterred by female speaker* `w10` *for cnn2 (top) and cnn3 (bottom).*

is consistent with speaker recognition studies [17]. The distinctive observation here is modeling of both F0 and formant information by a single CNN.

6. Conclusion

In this paper, we investigated an end-to-end gender identification approach using CNNs. We compared it against the approach of using cepstral coefficients and fundamental frequency to identify genders. Experimental studies on two corpora, namely, AVspoof and ASVspoof showed that the end-to-end approach consistently yields a better system, except in the case of acoustic mismatch condition. An analysis of the trained CNNs showed that depending upon the kernel width of the first convolution layer either formant information or both fundamental frequency and formant information is modeled by the CNNs for gender recognition. In recent years, features such as i-vectors have been used for gender recognition [10]. In [26], a comparison between the proposed CNN-based approach and i-vector based approach has been investigated for identifying gender under noisy conditions. It has been found that in both noisy condition training and denoised condition training the proposed CNN-based approach yields significantly better system.

7. Acknowledgment

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

8. References

[1] K. Wu and D. G. Childers, "Gender recognition from speech. part i: Coarse analysis," *The Journal of the Acoustical Society of America*, vol. 90, no. 4, pp. 1828–1840, 1991.

[2] D. G. Childers and K. Wu, "Gender recognition from speech. part ii: Fine analysis," *The Journal of the Acoustical Society of America*, vol. 90, no. 4, pp. 1841–1856, 1991.

[3] D. G. Childers, K. Wu, K. S. Bae, and D. M. Hicks, "Automatic recognition of gender by voice," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1988, pp. 603–606.

[4] E. S. Parris and M. J. Carey, "Language independent gender identification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 1996, pp. 685–688.

[5] M. Pronobis and M. Magimai.-Doss, "Analysis of f0 and cepstral features for robust automatic gender recognition," Idiap Research Report Idiap-RR-30-2009, 2009.

[6] S. Levitan, T. Mishra, and S. Bangalore, "Automatic identification of gender from speech," in *Proceedings of Speech Prosody*, 2016.

[7] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language, state-of-the-art and the challenge," *Computer Speech and Language*, 2012.

[8] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, 2013, special issue on Paralinguistics in Naturalistic Speech and Language.

[9] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Noth, "Age and gender recognition for telephone applications based on GMM supervectors and support vector machines," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 1605–1608.

[10] S. Ranjan, G. Liu, and J. H. Hansen, "An i-vector plda based gender identification approach for severely distorted and multilingual darpa rats data," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 331–337.

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

[12] D. Palaz, M. Magimai.-Doss, and R. Collobert, "End-to-end acoustic modeling using convolutional neural networks for automatic speech recognition," Idiap Research Report Idiap-RR-18-2016, 2016.

[13] 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.

[14] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[15] 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.

[16] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "End-to-end convolutional neural network-based voice presentation attack detection," in *IEEE IAPR International Joint Conference on Biometrics (IJCB)*, 2017.

[17] H. Muckenhirn, M. Magimai.-Doss, and S. Marcel, "Towards directly modeling raw speech signal for speaker verification using CNNs," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[18] S. Ergünay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *Proceedings of International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2015, pp. 1–6.

[19] W. Zhizheng, T. Kinnunen, N. W. D. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Proceedings of Interspeech*, 2015.

[20] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book*. Cambridge University Engineering Department, 2002.

[21] [Online]. Available: http://www.speech.kth.se/snack/

[22] [Online]. Available: http://www1.icsi.berkeley.edu/Speech/qn.html

[23] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A Matlab-like environment for machine learning," in *Proceedings of BigLearn, NIPS Workshop*, 2011.

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

[25] J. Hillenbrand, L. Getty, M. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *The Journal of the Acoustical Society of America*, vol. 97, pp. 3099–111, 06 1995.

[26] J. Sebastian, M. Kumar, D. S. P. Kumar, M. Magimai.-Doss, H. Murthy, and S. Narayanan, "Denoising and raw-waveform networks for weakly-supervised gender identification on noisy speech," in *Proceedings of Interspeech*, 2018.