Interspeech 2018
2-6 September 2018, Hyderabad

Denoising and Raw-waveform Networks for Weakly-Supervised Gender Identification on Noisy Speech

Jilt Sebastian[1,3,4], *Manoj Kumar*[2], *D. S. Pavan Kumar*[3,4], *Mathew Magimai.-Doss*[3], *Hema A. Murthy*[1] *and Shrikanth Narayanan*[2]

[1]Indian Institute of Technology Madras, India [2]University of Southern California, Los Angeles, USA
[3]Idiap Research Institute, Martigny, CH [4]École Polytechnique Fédérale de Lausanne, CH

jiltsebastian@gmail.com

Abstract

This paper presents a raw-waveform neural network and uses it along with a denoising network for clustering in weakly-supervised learning scenarios under extreme noise conditions. Specifically, we consider language independent Automatic Gender Recognition (AGR) on a set of varied noise conditions and Signal to Noise Ratios (SNRs). We formulate the denoising problem as a source separation task and train the system using a discriminative criterion in order to enhance output SNRs. A denoising Recurrent Neural Network (RNN) is first trained on a small subset (roughly one-fifth) of the data for learning a speech-specific mask. The denoised speech signal is then directly fed as input to a raw-waveform convolutional neural network (CNN) trained with denoised speech. We evaluate the standalone performance of denoiser in terms of various signal-to-noise measures and discuss its contribution towards robust AGR. An absolute improvement of 11.06% and 13.33% is achieved by the combined pipeline over the i-vector SVM baseline system for 0 dB and -5 dB SNR conditions, respectively. We further analyse the information captured by the first CNN layer in both noisy and denoised speech.

Index Terms: speech enhancement, Automatic Gender Recognition, convolutional neural network, recurrent neural network

1. Introduction

Weakly-supervised learning utilizes small amounts of training data, in contrast to fully supervised settings that rely on large amounts of training data (relative to test data). Such systems are particularly useful when it is possible to obtain only limited amounts of labeled data. Limited labeled data availability also challenges robust speech processing under unseen and noisy data conditions. It should be noted that most effective denoising methods in the state-of-the-art, however, are fully supervised in nature. Recent denoising algorithms use various types of neural networks for speech enhancement as opposed to traditional signal processing-based approaches. Several variants of Deep Neural Networks (DNNs) [1, 2] and Denoising Auto-Encoders (DAEs) [3] have been proposed for denoising the speech subject to non-stationary noise conditions. In this work, we present a denoising framework for low-resource speech interaction applications. In particular, we focus on the task of gender identification.

AGR from the speech signal is an essential preprocessing step for many applications and can prove to be challenging under weakly-supervised learning scenarios [4] or extreme noisy environments. Features derived from pitch and cepstral representations have been used in [5] and [6–8] under clean environments. Recent DNN-based gender classification systems employ transformed MFCCs as features [9]. Most of the approaches are restricted to the mono-lingual condition. Works such as [10–12] have however performed language-independent gender identification.

Gender identification has been performed on distorted speech in [11] using an i-vector PLDA system, on compressed speech in [12] using a combination of set of experts with neural network models. Language independent AGR is performed on noisy speech in [10] with Gaussian mixture models (GMMs). This model performs well on SNRs $\geq$0 dB. However, this work does not consider challenging noisy conditions, unseen language and noise conditions during test and, the results are reported at the utterance-level by considering all the vocalised segments together using Voice Activity Detector (VAD).

Raw-waveform methods have recently been proposed for various speech processing applications such as automatic speech recognition [13, 14], voice presentation attack detection [15], and emotion recognition [16] from speech. They are preferred due to their inherent ability to extract features specific to the application, and their superior performance. In a recent work, an end-to-end approach for gender classification, in similar lines of [13, 15, 17], has been developed [18]. It yielded better performance than standard acoustic features-based approach. We build on that work to develop a two-stage noise AGR system, where speech is denoised and then fed into the CNN for gender classification.

We perform language independent and weakly-supervised gender classification under challenging environmental noise conditions with unseen noise and language categories in the test set. We employ an SVM classifier as the baseline system, as it provides best AGR under weakly-supervised settings [4, 19]. It uses an i-vector based feature extractor. SVM is the popular choice for classification when only a limited amount of data is available for training [4]. The contributions of this work are two fold; First, we show that gender identification under highly-noisy conditions can be considerably improved using a denoising network. Second, we show that the raw-waveform CNN-based approach yields significantly better results than the i-vector based approach.

The rest of the paper is organized as below: Section 2 discusses the proposed pipeline for denoising and AGR. Section 3 describes the dataset and experimental procedure. Section 4 presents the results and analyses the performance. Conclusions are discussed in Section 5.

2. Methodology

Obtaining labeled data can be time-consuming, requires skilled personnel, and is also expensive. The natural alternative is to develop unsupervised or weakly-supervised models capable of handling variabilities on the test set. The latter may include

10.21437/Interspeech.2018-2321

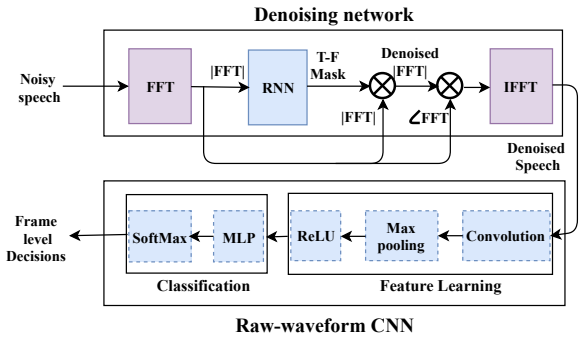

Figure 1: *Block schematic of the proposed approach.*

differences in speaker traits such as gender and age, linguistic capabilities, and environmental factors such as noise types (e.g. stationary/non-stationary, additive/convolutive) and noise levels. We propose to address some of these issues in our method. We use a small fraction of the data for training and validation, and the rest for testing. We simulate unseen noise and language conditions in our test set to investigate the robustness of the system to these conditions. Owing to these variabilities, it is vital to perform denoising as a preprocessing step. We propose to use a two-stage pipeline: speech denoising stage and subsequent gender identification stage (Figure 1).

2.1. Denoising stage

This stage consists of three components; feature extraction, time-frequency mask estimation using denoising network and speech reconstruction. The speech denoiser is inspired by speech separation models learning both the sources simultaneously [20]. This model learns *all* the sources of variability in its training. To account for highly-variable non-stationary noise and speech signal, we use a recurrent neural network (RNN) with magnitude spectrogram as its input. Magnitude spectrograms of the mixture of clean speech signal ($S[n,k]$, n and k are time and frequency indices, respectively) and noise signal ($N[n,k]$) are fed to the network. This work addresses denoising of additive noise. We formulate the separation problem as a classification problem to assign a soft label for speech and noise at each time-frequency bin [21]. An additional deterministic output layer is added to the network and it is jointly optimized with the normalized mask functions. A single model simultaneously learns the mask for both the speech and noise with a higher weight assigned to clean speech since it is fed into classification network. We minimize the Kullback Leibler divergence (KLD) objective [20]:

$$D(\widehat{y}_1[n]||y_1[n]) + D(\widehat{y}_2[n]||y_2[n]) \\ - \gamma(D(\widehat{y}_1[n]||y_2[n]) + D(\widehat{y}_2[n]||y_1[n]) \quad (1)$$

where , $\widehat{y}_i[n]$ and $y_i[n]$ represents the estimated and clean spectra respectively for source i, $D(X||Y)$ refers to KLD between X and Y. γ is an empirical parameter optimized to reduce the error between the original and estimated speech signals. The recurrent connection is given by,

$$\mathbf{h}(\mathbf{x_t}) = \mathbf{f}(\mathbf{W}\mathbf{h}(\mathbf{x_t}) + \mathbf{b} + \mathbf{V}\mathbf{h}(\mathbf{x_{t-1}})) \quad (2)$$

where, $\mathbf{W}$ and $\mathbf{V}$ are the weight matrices, $\mathbf{V}$ being the temporal weight matrix and $f(\cdot)$ a ReLU nonlinearity employed for separation [20]. We augment the training data by shifting either of the sources and mitigating the need for larger number of training samples. Denoised speech is obtained by multiplying the speech mask with the noisy magnitude spectrogram and using noisy phase (speech reconstruction in Figure 1).

The patterns associated with speech are added with various background noises which lead to variabilities in the spectrogram characteristics. Figure 2 shows the denoising process with an example taken from the test data [1]. Weakly-supervised classification is performed in the second stage by a raw-waveform CNN on the output of the denoiser (Figure 1). We postulate that the classifier trained with denoised output can provide better gender identification.

2.2. Raw-waveform CNN-based approach

Similar to [18], the network consists of two sub-stages: feature learning and classification. Feature learning consists of a 1-D convolutional layer with max pooling and ReLU non-linearities, which is repeated. Classification consists of a multilayer perceptron (MLP) with ReLU activations and a softmax output. The output layer performs the softmax operation to obtain frame-level gender posteriors. The decision is made by combining the frame-level posteriors. The feature stage and classification stage are jointly trained using stochastic gradient descent algorithm with cross entropy error criterion. In [18], it was found that for effective AGR, at least two convolution layers are needed. So we considered two architectures: (a) three convolution layers followed by one hidden layer, referred to as CNN1 and (b) two convolution layers followed by one hidden layer, referred to as CNN2.

Table 1 compares the architecture of CNN1 and CNN2. We use 300 ms window length ($\mathbf{W_{len}}$) with a 30 ms shift ($\mathbf{W_{shift}}$) for both architectures. In CNN1, the first convolution layer filter width is short, such that it models sub-segmental signal ($\approx$ 4 ms speech). We use CNN2 to examine the ability of raw-CNN methods with fewer parameters. This model has only $\approx$ 40% of the number of parameters compared to CNN1. CNN2 differs from CNN1 in the first convolution as it models "segmental" speech, i.e. about 20 ms speech ($\mathbf{N_{seq1}}$ = 150 samples). We use a max pooling size of 3 ($\mathbf{mp_i}, \mathbf{i} = \mathbf{1}..\mathbf{N}\ \forall$ convolutional layers $\mathbf{N}$). The third (final) convolutional layer of CNN1 has similar dimensions as the second layer.

3. Experiments

We perform denoising and language independent gender identification on the noisy version of CALLFRIEND corpus [2]. The noise signals are selected from various categories on publicly available DEMAND (Diverse Environments Multichannel Acoustic Noise Database) corpus [22]. The following subsections present details of the dataset, the experimental procedure, baseline system and the performance metrics.

3.1. Dataset

The CALLFRIEND corpus consists of unscripted two channel telephonic conversation between native speakers of 13 languages. We select audio from the train sets of Canadian French, Farsi, Hindi, Korean and German in this work. Data are pooled such that at least two speakers from each gender are selected per language. A total of 38 speakers (19 same-gender sessions) are selected and both sides of a conversation were added together to form a two-party, one-channel recording. It ensures that long silences are not present in the recording. This corresponds to a total of 582 sessions of five minutes each. The DEMAND

[1] mixture of "scafe" noise and female conversation in German

[2] https://catalog.ldc.upenn.edu/

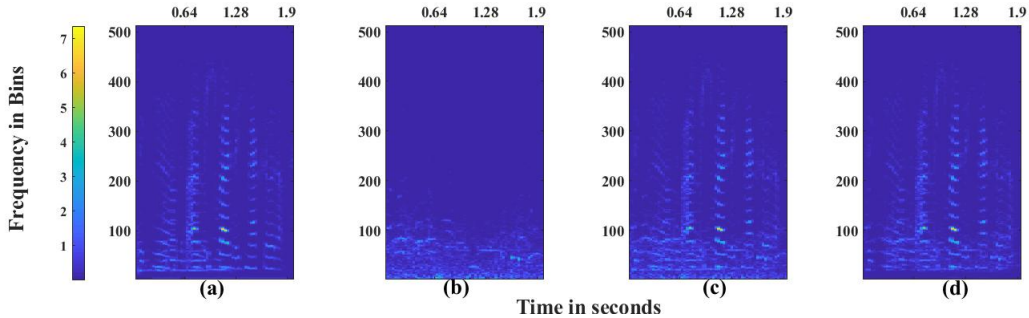

Figure 2: *An example of denoising process: Magnitude spectrograms of* **(a)** *clean speech,* **(b)** *noise signal,* **(c)** *noisy speech, and* **(d)** *denoised speech. Observe that the denoised speech is similar to the clean one.*

Table 1: *Comparison of two raw-waveform architectures.*

Parameters	CNN1	CNN2
number of conv. layers	3	2
L1 width/shift (in samples)/# filters	30/10/80	150/10/80
L2 width/shift (in frames)/# filters	7/1/60	7/1/60
Max pooling size/shift	3/1	3/1
number of hidden units	1024	100
Total number of parameters	433,114	184,042

dataset consists of five minutes, 16 channel (microphone distance between 5 cm and 21.8 cm) environment noise recordings for 18 different noise conditions, divided into six main categories (Domestic, Nature, Office, Public, Street and, Transportation). We select one condition from each category (*dliving, ooffice, omeeting, scafe, prestaurant, tbus*) to cover all kinds of environmental settings during the creation of noisy dataset for our experiments. We leave out one of the languages (German) and noise categories (meeting room noise) for the test set during both the denoising and gender classification part, to test the robustness of the system against unseen language and environmental conditions. The noises are mixed with the conversational speech at 0 dB and -5 dB SNRs.

3.2. Experimental procedure

We use windows of 128 ms and with 64 ms shift to compute the short-time Fourier transform. The RNN takes and predicts a 513 point spectrum with a previous time context. It consists of a feedforward hidden layer followed by a recurrent layer, each of 500 nodes with ReLU activations. The output layer is linear. 17% of the clean and noisy data is used (93 sessions) for its training (since it is weakly-supervised), that includes 4 conversation sessions for validation. We use 83% of the data for testing. The same denoiser trained with 0 dB SNR is used for evaluating the test segments under -5 dB SNR in order to analyse its robustness. Since we are interested in gender identification from the speech in adverse noise conditions, we only consider SNRs $\leq$0 dB.

The sessions are split into uniform segments of 2-second duration for classification. This is to ensure that the model is able to identify gender in with a short input signal. All possible combinations of noises and languages are considered with equal probability. A total of 84,240 such segments are used for the experiment. 30% of the dataset is used for training (21,494 segments) and cross-validation (3,582 segments), which includes all the training samples of the denoiser. The CNN is trained with an initial learning rate (LR) of 0.1. The LR is halved whenever the validation loss stagnates between successive epochs. Training is terminated when the LR drops below 10^{-6} and the final model is used for gender identification. The classifier is tested with 59,164 segments (70% of the dataset). We train both variants of the proposed classifier (CNN1, CNN2) with a different number of hyperparameters (Table 1). We use Keras [23] with TensorFlow [24] backend for building the raw-waveform CNNs.

3.3. Baseline system and performance metrics

We use an SVM classifier on i-vectors as our baseline method. SVMs are popularly chosen for learning from limited data [4]. I-vectors are used as feature representation for this task. The UBM-GMM with 2048 mixtures and 400 dimensional i-vector extractor are trained using 100 sessions from the AMI meeting corpus [25] down-sampled to 8 kHz. This method provides state-of-the-art gender identification system for weakly-supervised learning [4]. We use SVM classifier with Radial Basis Function kernel and the model is trained using scikit-learn python package [26]. We analyze the effect of denoiser on the baseline as well.

Table 2: *Denoiser performance at different noise levels.*

Measure	**Binary Mask**		**Soft Mask**	
	-5 dB	**0 dB**	**-5 dB**	**0 dB**
GNSDR	18.09	11.12	18.24	19.50
GSIR	23.57	20.30	20.03	19.50
GSAR	14.28	13.05	14.66	14.15

Since the denoiser output is used for further processing, its metrics should be able to accommodate both the amount of noise and artifacts introduced in the denoising process as opposed to the traditional evaluation metrics (SNR and PESQ). Signal to Interference Ratio (SIR), Signal to Artifacts Ratio (SAR) and Signal to Distortion Ratio (SDR) from BSS Evaluation Metrics [27] are chosen for evaluation. SIR refers to the amount of noise contained in the separated signal (equivalent to SNR), SAR denotes the amount of artifacts introduced after separation and SDR represent the overall separation quality. We report the improvement in SDR with respect to the mixture in terms of Normalized SDR (NSDR). Each segment is weighted by its length and averaged across total number of segments to obtain Global measures (GNSDR, GSIR and, GSAR). Higher the values, better is the separation quality. We report unweighted average recall (UAR) for gender classification since it is robust to class imbalance.

4. Results and Discussions

We report the performance of denoiser in Table 2. We use both binary and soft masks in our experiments and observe that binary

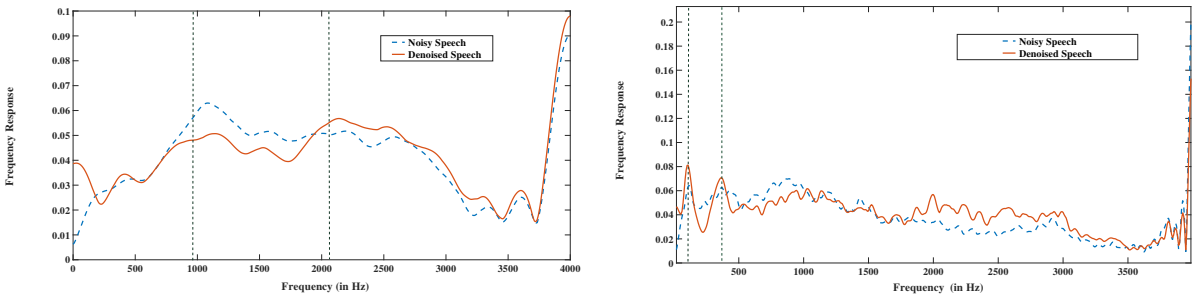

Figure 3: *Cumulative Frequency Response of Layer1 filters in CNN1* **(left)** *and CNN2* **(right)**.

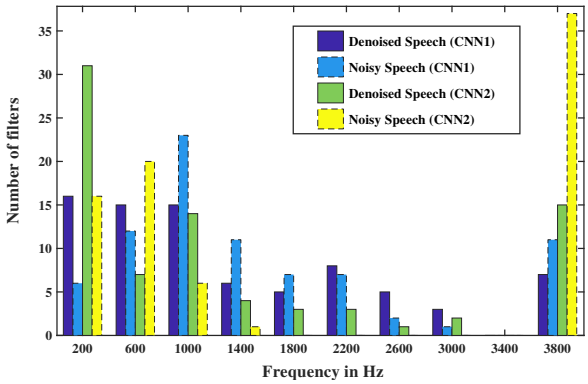

Figure 4: *Histograms of peak frequency responses of filters in sorted order.*

mask performs better compared to soft mask in general. The denoiser is trained at 0 dB mixing condition and tested on both 0 dB and -5 dB conditions. Discriminative training causes larger SIR values [20]. Denoiser performs equally well for unseen noise category (*omeeting*), language (*German*) and their combinations [3]. The denoiser has all evaluated metrics above 10 dB. We further analyze its role in gender classification.

The results of language independent AGR are shown in Table 3. Systems trained with noisy speech at 0 dB SNR and its denoised version are used directly for testing the noisy speech at -5 dB SNR and its denoised version respectively. All systems show a consistent improvement over the baseline under both noise levels. Raw-waveform CNN architectures perform significantly better than the baseline. As expected, the UAR is higher for 0 dB as compared to -5 dB mixing condition across the architectures. Denoiser improves the performance of *all* of them. An absolute improvement of 11.06% and 13.33% is achieved by a combination of denoiser and raw CNN method over the baseline for 0 and -5 dB SNRs respectively.

Table 3: *Gender identification performance in terms of UAR (%) at different noise levels.*

System	Noisy		Denoised	
	-5 dB	0 dB	-5 dB	0 dB
Baseline	76.84	81.95	79.83	83.34
CNN2	83.86	89.13	88.00	91.53
CNN1	87.47	91.31	**90.17**	**93.01**

We plot Cumulative Frequency Response (CFR) of the learned filters in Figure 3. It is obtained by normalizing the sum of all the filter responses [17] and shows the frequency regions the filters emphasize collectively. The filters learned from the noisy speech and denoised speech are *similar*, except that the denoised versions provide room for a clearer analysis. CNN1 seems to give emphasis to formant regions, around 1000 and 2000Hz whereas, CNN2 captures gender discriminative information in low frequency regions as well as high frequency regions. Specifically, CNN2 CFR has two peaks at 101 Hz and 351 Hz, potentially modeling male and female average fundamental frequency respectively. These observations indicate that CNN with different architectures learns to weigh the frequency spectrum at different resolutions - capturing vocal tract information in one (CNN1) and fundamental frequency in another (CNN2).

We plot histograms of peak frequency responses of filters in sorted order in Figure 4. There are larger number of peak frequency filters in low-frequency region ($\leq$ 400 Hz) for the denoised speech as compared to noisy speech, possibly due to pitch and low-frequency formants easier to learn in denoised conditions. The plot also reveals the frequency selective nature of individual filters. Observe that, there are more number of filters with peak frequency in the high-frequency region in noisy speech than that in denoised speech. This could be due to the artifacts and presence of high-frequency noise. Further analyses inclusive of original clean speech models are required to understand them.

5. Conclusions

We presented a two-stage pipeline for AGR under noisy conditions. In this pipeline, speech signal is denoised using an RNN and fed to gender classification system. We investigated two types of systems; i-vector based SVM system and, CNN-based end-to-end system. Experimental studies show that, irrespective of the type of AGR system, RNN-based denoising improves the classification performance. A comparison across AGR systems showed that the CNN-based approach outperforms the i-vector based SVM approach under all noise levels for both noisy condition training and denoised condition training. This shows that joint learning of feature and classifier from raw speech signal is beneficial for noise-robust AGR.

6. Acknowledgements

This work was partially supported by Swiss Government Excellence Scholarship Project with ESKAS No: 2017.0575, Simons Foundation, and HASLER Foundation project FLOSS.

7. References

[1] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2014.

[3] Performance and denoised examples are available at:`https://sites.google.com/site/weaklysupervisedgenderid/`

[2] Xu, Yong, Du, Jun, Dai, Li-Rong, and Lee, Chin-Hui, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[3] X. Feng, Y. Zhang, and J. Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1759–1763.

[4] J. Ahmad, M. Fiaz, S.-i. Kwon, M. Sodanil, B. Vo, and S. W. Baik, "Gender identification using mfcc for telephone applications-a comparative study," *arXiv preprint arXiv:1601.01577*, 2016.

[5] B. D. Barkana and J. Zhou, "A new pitch-range based feature set for a speaker's age and gender classification," *Applied Acoustics*, vol. 98, pp. 52–61, 2015.

[6] 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.

[7] K.-H. Lee, S.-I. Kang, D.-H. Kim, and J.-H. Chang, "A support vector machine-based gender identification using speech signal," *IEICE transactions on communications*, vol. 91, no. 10, pp. 3326–3329, 2008.

[8] E. Ramdinmawii and V. Mittal, "Gender identification from speech signal by examining the speech production characteristics," in *Signal Processing and Communication (ICSC), 2016 International Conference on*. IEEE, 2016, pp. 244–249.

[9] Z. Qawaqneh, A. A. Mallouh, and B. D. Barkana, "Deep neural network framework and transformed mfccs for speaker's age and gender classification," *Knowledge-Based Systems*, vol. 115, pp. 5–14, 2017.

[10] Y.-M. Zeng, Z.-Y. Wu, T. Falk, and W.-Y. Chan, "Robust gmm based gender classification using pitch and rasta-plp parameters of speech," in *Machine Learning and Cybernetics, 2006 International Conference on*. IEEE, 2006, pp. 3376–3379.

[11] 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.

[12] H. Harb and L. Chen, "Gender identification using a general audio classifier," in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, vol. 2. IEEE, 2003, pp. II–733.

[13] D. Palaz, R. Collobert, and M. M. Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," *arXiv preprint arXiv:1304.1018*, 2013.

[14] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, and M. Bacchiani, "Factored spatial and spectral multichannel raw waveform cldnns," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5075–5079.

[15] 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.

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

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

[18] S. H. Kabil, H. Muckenhirn, and M. Magimai.-Doss, "On learning to identify genders from raw speech signal using CNNs," in *INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association, September 2-6, Hyderabad, India, Proceedings*, 2018.

[19] J. Ahmad, K. Muhammad, S.-i. Kwon, S. W. Baik, and S. Rho, "Dempster-shafer fusion based gender recognition for speech analysis applications," in *Platform Technology and Service (PlatCon), 2016 International Conference on*. IEEE, 2016, pp. 1–4.

[20] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2136–2147, 2015.

[21] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1849–1858, 2014.

[22] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, no. 1. ASA, 2013, p. 035081.

[23] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[24] M. A. et. al., "TensorFlow: Large-scale machine learning on heterogeneous systems," software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[25] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The ami meeting corpus: A pre-announcement," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 28–39.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[27] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.