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

Jilt Sebastian1,3,4, Manoj Kumar2, D. S. Pavan Kumar3,4, Mathew Magimai-Doss3, Hema A. Murthy1 and Shrikanth Narayanan2

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

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)12 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 or extreme noisy environments. Features derived from pitch and cepstral representations have been used in5 and 6 under clean environments. Recent DNN-based gender classification systems employ transformed MFCCs as features. Most of the approaches are restricted to the mono-lingual condition. Works such as10–12 have however performed language-independent gender identification.

Gender identification has been performed on distorted speech in17 using an i-vector PLDA system, on compressed speech in12 using a combination of set of experts with neural network models. Language independent AGR is performed on noisy speech in10 with Gaussian mixture models (GMMs). This model performs well on SNRs ≥ 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 recognition13,4, voice presentation attack detection15, and emotion recognition16 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 of13,15,17, has been developed18. 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 settings4,19. 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 training4. 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: Section2 discusses the proposed pipeline for denoising and AGR. Section3 describes the dataset and experimental procedure. Section4 presents the results and analyses the performance. Conclusions are discussed in Section5.

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
We propose to address some of these issues in our method. We weight matrix and where,

\[ \hat{y} \]

To account for highly-variable non-stationary noise and speech level, we use a recurrent neural network (RNN) with magnitude spectrogram as its input. Magnitude spectrograms of the mixture of clean speech signal (\( S[n, k] \)) and noise signal (\( N[n, k] \)) are fed to the network. This work addresses denoising of additive noise. We minimize a small fraction of the data for training and validation, and this 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]:

\[
D(\tilde{y}_i[n]|y_i[n]) + D(\tilde{y}_2[n]|y_2[n]) \\
- \gamma(D(\tilde{y}_1[n]|y_2[n]) + D(\tilde{y}_2[n]|y_1[n]))
\]

(1)

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

\[
h(x_t) = f(Wh(x_t) + b + \text{ReLU}(h(x_{t-1})))
\]

(2)

where, \( W \) and \( V \) are the weight matrices, \( V \) 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. 

### 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 (\( W_{\text{len}} \)) with a 30 ms shift (\( W_{\text{shift}} \)) 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 (\( N_{\text{seg}} = 150 \) samples). We use a max pooling size of 3 (\( m_p \), \( i = 1, N \forall \) convolutional layers \( N \)). 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/
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 (living, office, meeting, café, restaurant, bus) 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 ≤ 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.

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

<table>
<thead>
<tr>
<th>System</th>
<th>Noisy -5 dB</th>
<th>Noisy 0 dB</th>
<th>Denoised -5 dB</th>
<th>Denoised 0 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>76.84</td>
<td>81.95</td>
<td>79.83</td>
<td>83.34</td>
</tr>
<tr>
<td>CNN2</td>
<td>83.86</td>
<td>89.13</td>
<td>88.00</td>
<td>91.53</td>
</tr>
<tr>
<td>CNN1</td>
<td>87.47</td>
<td>91.31</td>
<td>90.17</td>
<td>93.01</td>
</tr>
</tbody>
</table>

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 weight the frequency spectrum at different resolutions - capturing vocal tract information in one (CNN1) and fundamental frequency in another (CNN2).

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.

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


