Speaker Gender Classification based on an Improved Deep Learning Approach

Mohamed Anouar Ben messaoud[1,2], *Aicha Bouzid*[1]

[1]*LR11ES17 Signal, Image and Information Technology Laboratory, National School of Engineers of Tunis, University of Tunis El Manar*
[2]*Faculty of Sciences of Tunis, University of Tunis El Manar*
Tunis, Tunisia
anouar.benmessaoud@fst.utm.tn,bouzidacha@yahoo.fr

Kurzfassung: With the great evolution of technology, Speaker gender and age classification is one of the major problems for large range of applications in speech analysis and recognition. The identification of speakers has become crucial in the cases of criminal suspect, speech recognition, speech emotion, and computer-aided physiological. To improve the accuracy of speaker gender classification, we must generate robust features with a depth classifier. With the promising results giving by machine learning for classification problem, our approach has taken advantage of deep learning. In this paper, we propose to apply a speaker gender classification based on the Recurrent Neural Network (RNN) which is able to determine the long term dependencies of a sequential speech signal. The most popular RNN is Long Short-Term Memory (LSTM) model. However, it has a complex design which makes it difficult to implement. So, we refine the LSTM model to our proposed Simplified Gated Recurrent Units (SGRUs) by using an efficient architecture with only two multiplicative gates more suitable for speech classification. Our approach is decomposed into two essential steps. First, we generate the features from our model train based on SGRUs by removing of reset gates to limit redundancy and reduce the number of parameters without affect the system performance. Second, we use the Rectified Linear Units (ReLU) activations to learn long-term dependencies without slow down the training process. In Our architecture, we modify the level of dropout and increase the depth of the network. The architecture was tried on a public challenging database. Experiment results show that our approach presenting a high accuracy surpassing other recent methods of gender classification task.

1 Introduction

In speech processing, we need information about speaker such as gender, accent, language, speaker characteristics or age to improve speech analysis [1, 2].

The voice gender classification system allows to identify the speaker is male or female so that be used for example in health related communications to call forwarding, in automatic speech recognition (ASR) to limit the search domain to speech signals from the same gender, in multimedia to use the speaker's gender as a label for the indexed speech signals, or in audio compression when gender identification dependent speech coders must be applied.

In literature, we can found three categories of gender identification: gender dependent models [3], gender independent models [4], and gender based on speech coders [5].

With the application of advanced machine learning approaches, many works could achieve to classify correctly short utterances [6] and it's a solution to abstract a task. It's differ by types of features applied as an input to the neural network architecture. For example, in [7], the authors used the linear predictive coding coefficients (LPC) to consider a multi-layer perceptron as a voice gender classifier. In [8], the authors proposed the energy entropy and zero crossing rates to feed the neural network combined with fuzzy logic system, and in the work of Sas, two parameters mel-frequency cepstral coefficients (MFCC) and pitch estimation are used as features to design the neural network architecture [9].

In order to classify these features, we can employed several classifiers like linear regression, logistic regression, *K*-nearest neighbour (KNN) [10], support vector machine (SVM) [11], random forests, and convolutional neural network (CNN) [12].

In the last decade, many tasks have been successfully realized using deep neural networks. In [13], Bisio et al. proposed an approach based on multiple unsupervised SVM with a dynamic training for gender identification. In [11], Pahwa et al. have applied an SVM combined with MFCC as the feature inputs to a neural network system using a stacking technique. Pribil et al. have proposed Gaussian Mixture Model (GMM) classifier with two levels [14]. Levitan et al. have trained the random forest classifier on MFCC and fundamental frequency features [15]. The approach proposed by Buyukyilmaz et al. [16] is based on multilayer perceptron deep learning model to gender Recognition. In [17], the authors have used gradient boosting machines algorithm (GBMs) for gender identification.

In this paper, we use the MFCC and fundamental frequency of the speaker as input features to our recurrent neural network (RNN). Our deep neural network model of voice gender identification is decomposed into two stages. First, we remove the reset gates. Second, we apply the rectified linear units (ReLU) activations to our Simplified Gated Recurrent Units (SGRUs). With this SGRUs architecture, our system is training. Then, the test data are classified to obtain the gender voice of the speaker as output of our system.

2 Proposed Approach for Gender Classification

In this paper, we propose an approach based on simplified gated recurrent units (SGRUs) architecture. Also, we demonstrate that when our architecture uses just two multiplicative gates is more suitable for gender classification. Our architecure is decomposed on two steps. In the first step, we remove the reset gate. However, in the second step, we apply rectified linear units activations and we add a batch normalisation.

In figure 1, we illustrate our proposed approach. We can resume by the training of the features, followed by a classification system.

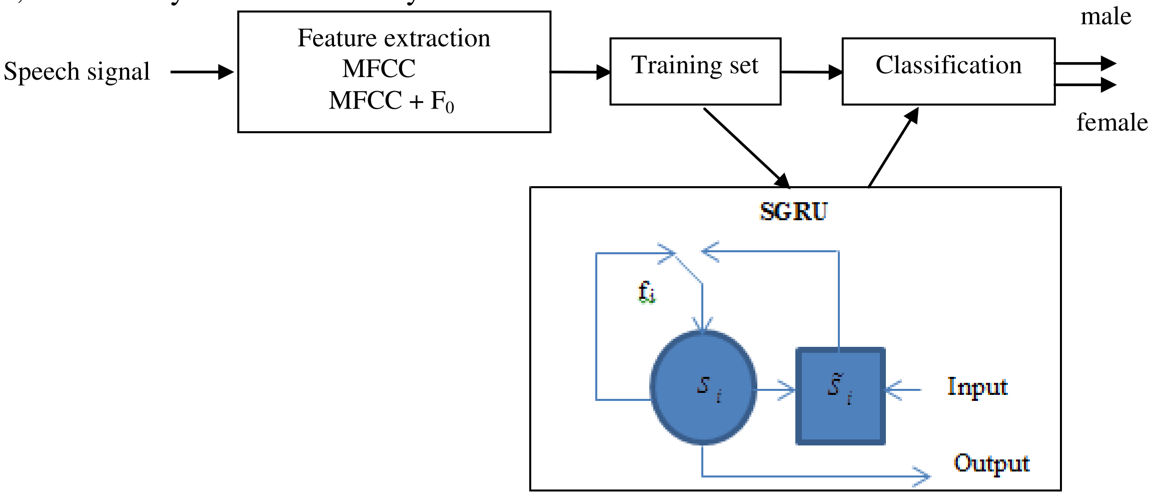

Figure 1 – Overview of the application of our Gated Recurrent Architecture

2.1 Simplified GRU

Standard Gated Recurrent Unit architecture is described by the reset b_i and update f_i gates. It defined by the following equation:

$$\begin{cases} b_i = \lambda\left(C_b x_i + R_b s_{i-1} + w_b\right) \\ f_i = \lambda\left(C_f x_i + R_f s_{i-1} + w_f\right) \end{cases} \quad (1)$$

Our network is fed by the speech vector x_i with its features. For this model, we use R_b, R_f as the recurrent weights and C_b, C_f as the feed-forward connections matrices. For each current time i, s_i represents the state vector. The vectors w_b, and w_f are used as trainable bias. To ensure that the two gate vectors take values between 0 and 1, we apply activations in the form of sigmoid functions $\lambda(.)$.

We determine the weighting factors by the update gate f_i that give the decision about the update of activations by their units and it is used as a key component for learning long-term dependencies. So, if the update gate f_i is close to zero then our network applies the candidate state with rectified linear units that depend on the closer hidden states and the current input. Else, if it is equal to one then the previous state is kept unchanged.

According to the semantic relation of the speech, we notice a redundancy occurs in the activations of update and reset gates. When we give more importance to the current data, we will have small b_i values from the GRU model. Also, we get the small values of f_i when we use only the update gate. This solution will weigh down the candidate state. This candidate depends on the most recent history as well as the current input. Also, a high value can be given to reset and update gates whose objective is to give more importance to past states. For a standard GRU, we can notice a temporal correlation in the average activations of f_i and b_i gates.

Based on these observations, we have proposed to ignore the stored memory and avoid a biased decision by an uncorrelated history. So the reset gate b_i might not be useful. In our architecture called Simplified Gated Recurrent Unit (SGRU), we remove the reset unit to obtain a single gate model in order to limit redundancy in the gating mechanism. After the use of a single gate, we observe an efficient reduction in the computational complexity by reducing the number of parameters which allows reaching the performance of a standard GRU.

2.2 Modified activations functions

In the second stage, we use our Rectified Linear Units (ReLU) activations instead of the standard hyperbolic tangent based on the fact that the standard hyperbolic tangent (tanh) provokes vanishing gradient and degrades the training process. Then, we replace them with neurons based on Rectified Linear Units in order to reduce these limitations and to learn long-term dependencies without slow down the training process. For this, you just need an appropriate orthogonal initialization in order to form stable ReLU functions and eliminate the problem originated by the unbounded ReLU functions.

After replacing the tanh function with the Rectified Linear Units (ReLU), we obtain the following equations:

$$\begin{cases} \tilde{s}_i = \mathrm{ReLU}\left(C_s x_i + R_s s_{i-1} + w_s\right) \\ s_i = f_i s_{i-1} + \left(1 - f_i\right)\tilde{s}_i \end{cases} \quad (2)$$

Finally, we normalize the variance and the mean for each layer pre-activation and apply it to recurrent connections.

Such a technique has shown to be important to avoid the numerical issues, to speed-up the training procedure, and to improve the system performance.

The proposed normalization technique allows us to resize the neuron pre-activations and subsequently bounding the values of the ReLU neurons so that RNN benefits of such activations. We initialize the gain factor of the proposed normalization at a value equal to 0.1.

3 Experiments and Results

3.1 Dataset

To evaluate the proposed architecture, a set of experiments was performed with SITW database [18]. The Speakers in the Wild (SITW) speaker recognition database includes recordings of 299 speakers, with eight sessions per speaker. The SITW is composed of reverberation, compression artifacts, and real noise.

3.2 System setup

For our evaluation, we apply an architecture based on multiple bidirectional recurrent layers obtained by concatenating the forward with backward hidden states followed by the softmax context-dependent classifier. We apply an orthogonal and faster convergence initialization to initialise recurrent weights and the feed-forward connections of the architecture, respectively.

Mini-batches of six sorted sentences were processing by the training stage to make the stability of gradients and improve the performance. Also, we apply the Adam technique for eighteen epochs to obtain an optimization without gradient truncation to ensure arbitrary learning of long time dependencies.

As main hyper parameters of the model, we apply an initial learning rate of $12*10^{-4}$ with two hidden layers and about two hundred neurons. Also, we consider 70% of dataset were used as training set, and the rest as test set in the same corpus.

3.3 Results

We achieve our evaluation with two set of features. In the first set, we just use 38 MFFCCs input features. We calculate the feature vectors every 10 ms. However, in the second set, we apply a set of feature based on MFCCs and fundamental frequency (F_0) features such as we compute F_0 in a frame size of 25.6 ms. We run the classification stage 1000 times with randomizing training to evaluate the accuracy.

In table 1, we present the results of accuracy of our proposed approach (SGRUs) compared to the traditional recurrent neural network (RNN) based on GRU with reset gates and tanh activations, and long short-term memory (LSTM).

Table 1 - Accuracy Results

System	Accuracy (%)	
	Set Features (MFCCs)	Set Features (MFCCs + F_0)
SGRU	94.6	97.5
RNN	89.5	91.8
LSTM	90.2	92.3

Table 1 shows that our approach achieved improvement in accuracy. We found that removing the reset gate does not affect the performance of system. In addition, the application of ReLU functions has allowed removing the vanishing gradient problem.

We can observe that our SGRUs outperforms the LSTM and RNN based on standard GRU architectures. The results show that the ignorance of reset gate and the application of ReLU activations increase the performance.

In table 2, we compared our approach to three state-of-the art algorithms of gender classification.

Table 2 - Comparison the state-of-the art systems with the proposed system

System	**Accuracy (%)**	
	Set Features (MFCCs)	**Set Features (MFCCs + F_0)**
Our system	94.6	97.5
CNN [12]	91.5	93.1
KNN [10]	87.2	88.3

We can observe that our SGRUs approach outperforms the convolutional neural network (CNN) and *K*-nearest neighbors (KNN) architectures.

We note that for the two set of features, our proposed approach architecture gives the best results. This comparison shows the effectiveness and the robustness of our approach.

According to the two experiences, we can remark that the second set of feature based on MFCCs and F_0 features give more performance. So the system that models the two types (MFCCs+ F_0) identifies better the difference between genders.

4 Conclusion

In this paper, we presented a gender identification approach based on simplified Gated Recurrent Units architecture. In our approach, we removed the reset gate, and we applied ReLU activations. Experimental studies on SITW corpus showed that our approach yields the better results with a set of features composed by MFCCs and F_0. We obtain a high accuracy compared to state-of-the-art results, and it is easy to implement our approach due to their arithmetic simplicity. The future work will be about an extension of this approach to noisy environment.

5 Literatur

[1] Ben Messaoud, M.A., und A. Bouzid: *Sparse representations for single channel speech enhancement based on voiced/unvoiced classification. Journal of Circuits, Systems, and Signal Processing (CSSP), Springer,* Vol. 36, No. 5, S. 1912 - 1933, 2017.

[2] Ben Messaoud, M.A., und A. Bouzid: *A New biologically inspired fuzzy expert system-based voiced/unvoiced decision algorithm for speech enhancement. Journal of Cognitive computation, Springer*, Vol. 8, No. 3, S. 478 - 493, 2016.

[3] Potamitis, I., N. Fakotakis, und G. Kokkinakis: *Gender-dependent and speaker dependent speech enhancement.* In *Proc. IEEE International Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, S. 249 – 252, USA, 2002.

[4] Acero, A., und X. Huang: *Speaker and gender normalization for HMM.* In *Proc. IEEE International Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, S. 342 – 345, USA, 1996.

[5] Marston, D.: *Gender adapted speech coding*. In *Proc. IEEE International Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, USA, 1998.

[6] Levitan, S.I., T. Mishra, und S. Bangalore. *Automatic identification of gender from speech.* In *Speech Prosody*, USA, 2016.

[7] Konig Y., und N. Morgan: *GDNN: A gender-dependent N.N. for continuous speech recognition.* In *Proc. International Joint Conference on Neural Networks (IJCNN)*, USA, 1992.

[8] MEENA, K., K. SUBRAMANIAM, und M. GOMATHY: *Gender classification in speech recognition using fuzzy logic and NN. International Arabic Journal Information Technology.*, Vol. 10, No. 5, 2013.

[9] Sas J., und A. Sas: *Gender recognition using neural networks and automatic speech recognition techniques. Journal of Medical Informatics & Technologies*, Vol. 22, S. 1642 – 6037, 2013.

[10] SOLER, J., und L. WANNER: *A Semi-supervised approach for gender identification.* In *Proc. International Joint Conference on Language Resources and Evaluation,* S. 1282-1287, Slovenea, 2016.

[11] PAHWA, A., und G. AGGARWAL: Speech feature extraction for gender recognition. International Journal Image, Graphics, and Signal Processing, Vol. 9, S. 17 - 25,2016.

[12] KABIL, S.H., H. MUCKENHIRN, und M. MAGIMAI: *On Learning to identify genders from raw speech signal using CNNs*. In *Proc. Interspeech*, 2018.

[13] BISIO, I., F. LAVAGETTO, M. MARCHESE, A. SCIARRONE, und C. FRA, M. VALLA: Spectra: A speech processing platform as smartphone application. In *Proc. of IEEE International Conference on Communications (ICC)*, S. 7030 – 7035, UK, 2015.

[14] PRIBIL, J, A. PRIBILOVA, und J. MATOUSEK: *GMM-based speaker gender and age classification after voice conversion.* In *Proc. International Workshop on SPLINE*, Denmark, S. 1 – 5, 2016.

[15] Holzinger, A.: *Introduction to machine learning and knowledge extraction (MAKE). Machine Learning & Knowledge Extraction.* Vol. 1, S. 1 – 20, 2017.

[16] BUYUKYILMAZ, M., und A.O. CIBIKDIKEN: *Voice Gender Recognition Using Deep Learning*. In *Proc. International Conference on Modeling, Simulation and Optimization Technolgies and Applications (MSOTA)*. Vol. 58, S. 409 – 411, 2016.

[17] ZVAREVASHE, K., und O.O. OLUGBARA : *Gender voice recognition using random forest recursive feature elimination with gradient boosting machines.* In *Proc. International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, South Africa, S. 1 – 6, 2018.

[18] MCLAREN, M., L. FERRER, D. CASTAN, und A. LAWSON: *The Speakers in the Wild (SITW) Speaker Recognition Database*. In *Proc. Interspeech*, S. 818 – 822, 2016.