Multimedia Tools and Applications (2022) 81:3535–3552
https://doi.org/10.1007/s11042-021-11614-4

1188: ARTIFICIAL INTELLIGENCE FOR PHYSICAL AGENTS

Age group classification and gender recognition from speech with temporal convolutional neural networks

Héctor A. Sánchez-Hevia[1] · Roberto Gil-Pita[1] · Manuel Utrilla-Manso[1] · Manuel Rosa-Zurera[1]

Received: 31 January 2021 / Revised: 27 May 2021 / Accepted: 22 September 2021 /
Published online: 13 January 2022

Abstract

This paper analyses the performance of different types of Deep Neural Networks to jointly estimate age and identify gender from speech, to be applied in Interactive Voice Response systems available in call centres. Deep Neural Networks are used, because they have recently demonstrated discriminative and representation capabilities in a wide range of applications, including speech processing problems based on feature extraction and selection. Networks with different sizes are analysed to obtain information on how performance depends on the network architecture and the number of free parameters. The speech corpus used for the experiments is Mozilla's Common Voice dataset, an open and crowdsourced speech corpus. The results are really good for gender classification, independently of the type of neural network, but improve with the network size. Regarding the classification by age groups, the combination of convolutional neural networks and temporal neural networks seems to be the best option among the analysed, and again, the larger the size of the network, the better the results. The results are promising for use in IVR systems, with the best systems achieving a gender identification error of less than 2% and a classification error by age group of less than 20%.

Keywords Interactive voice response · Age estimation · Gender recognition · Human-robot interaction · Machine learning

This work has been partially funded by the Spanish Ministry of Economy, Industry and Competitiveness with project RTC-2016-4687-7, the Spanish Ministry of Science, Innovation and Universities, with project RTI2018-098085-B-C42 (MSIU/FEDER), and by the Community of Madrid and University of Alcala under project EPU-INV/2020/003.

✉ Manuel Rosa-Zurera
manuel.rosa@uah.es

Héctor A. Sánchez-Hevia
hectoradrian.sanchez@uah.es

Roberto Gil-Pita
roberto.gil@uah.es

Manuel Utrilla-Manso
manuel.utrilla@uah.es

[1] Signal Theory and Communications Department, University of Alcalá, Madrid, Spain

1 Introduction

This paper deals with the design of algorithms to classify speakers into age and gender groups, to be applied in call centres with 'Interactive Voice Response' (IVR) systems. It is a technology that allows a computer to interact with humans through the use of voice, and therefore, IVR is an example of system for human-machine interaction, which automatically answers, processes, and routes customer calls to the most appropriate operator in contact centres. IVRs are Spoken Language Understanding (SLU) systems based on the How-May-I-Help-You (HMIHY) task [12].

IVRs began to be used commercially in the 70s by the banking system, with the aim of offering customer account balances. At the beginning, they were very closed applications with very high costs. In the following years, technology developed exponentially, achieving much more reliability, and adding functionalities such as speech recognition, text-to-speech conversion, fax capabilities, and Internet integration. From its inception, the main milestones in the evolution of IVRs are the following:

- The first systems allowed the selection of options by the user through the telephone keypad, as a way of answering the questions posed by voice. These systems were not very intuitive, since the options were grouped according to the internal organisation of the company, and without taking into account the needs of the users.
- With the arrival of speech recognition technologies, IVRs were developed that could distinguish keywords such as 'yes', no', 'balance', 'invoice', 'numbers', etc. In this way, the user was not forced to press the digits on the telephone keypad, but could instead pronounce the commands or keywords.
- The evolution of automatic speech recognition (ASR) systems giving solution to the problem of natural language understanding, has allowed the user to utter more complex phrases, such as 'I want my last invoice', 'recharge the balance', etc. At the beginning, the user was presented with a closed set of options designed a priori, which forced him to adjust the real reason for the call to one of the defined themes. It is from the incorporation of linguistic recognition, which will allow the 'Open Question' option, when the user has possibility of freely expressing the reason for the call.

Call centres agents serve customers through phone calls. Most of the costs are spent on human resources, so a great effort has been made to optimise the use of agents, considering they are not homogenous, and have different experience and skills, handling customers requests with different speeds [40].

For this purpose, different routing strategies are applied in IVR systems, like direct routing, self-service routing, skill-based routing, or data-directed routing. Direct routing allows callers to select their preferences or needs among a group of possibilities, and are routed to the selected department or employee. Self-service routing allows callers to select their preferences or needs and receive information with prerecorded messages. Skill-based routing distributes incoming calls to the most appropriate operator, rather than simply choosing the next available agent. Data-directed routing is based on predicting which agent from available agents would be best suited for better customer communication. The prediction uses historical data from agents and customers, or other information that could be obtained from the customer voice, such as emotions, age and gender. Both information extraction and best match prediction can be implemented with machine learning techniques [20].

The basic scheme of an IVR system is represented in Fig. 1. Currently available IVR systems perform automatic caller intent classification using continuous speech recognition, which translates speech into text, and semantic text classification, which classifies speech into a predefined call intent [5]. These systems also extract biometric and emotional information from speech, which complements the lexical and syntactic information, so that human-machine interaction is similar to human-human interaction. The complete set of information is used to feed the algorithm which implements the routing strategy.

In this paper, we focus on data-directed routing, studying the possibility of implementing machine learning systems to obtain biometric data from the customer's voice with good enough results to be used in IVR systems. Speech contains information about the customer's age, gender, and emotional state. Information about emotions can be obtained directly from the speech signal, but also from the text generated by using speech recognition systems. Information about the acoustic environment is also relevant. In conclusion, IVR systems must include complex signal processing and natural language processing algorithms:

- Speech recognition, to generate text from speech, complemented by natural language processing to extract its meaning.
- Speech processing to extract information about emotions, biometric information (gender, age, etc.) to improve routing effectiveness.
- Acoustic environment monitoring, to obtain information about the environment where the user is and relate that information to the meaning of the message and the speaker's emotions.

Interest in emotion detection to be applied in call centres is not new. The first studies date back to about fifteen years, when the first experiments with real-life signals were carried out [39], and the results obtained with lexical and paralinguistic cues were compared [10]. The most interesting emotion from an application point of view in call centres is anger, which has recently received attention in [27] and [23].

The main application of age and gender recognition in this context is the performance of automated demographic analysis, which can be used to improve market research or personalise the experience for different target audiences, among other uses. Gender, age, and accent have a notable impact on confidence in the information provided and on satisfaction with the care provided by agents [9][24].

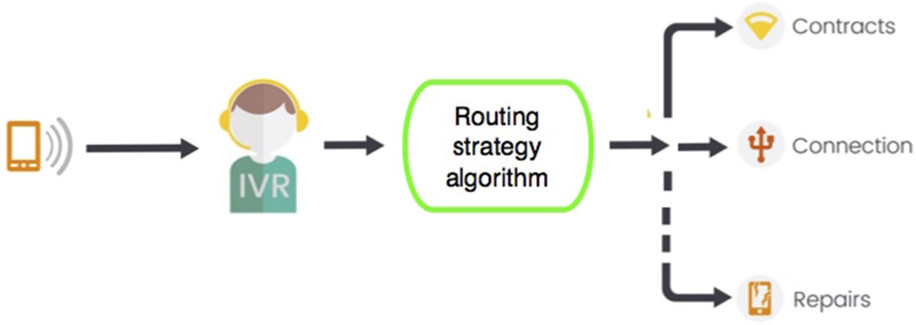

Fig. 1 Basic scheme of an IVR system

Age has been estimated using images [14] or from auditory information [15], using different types of learning machines. The speech signal also contains age information and can be used for this purpose. Age estimation from speech recordings (viewed as a regression problem) has been approached several times in the last two decades. In [22], speech signals sampled at 16kHz, and PCM coded with 16 bits were used to calculate MFCC, ΔMFCC and Power coefficients, to be applied to Linear Discriminant Analysers (LDA), and Neural Networks to estimate speakers' age. The dataset was divided into two groups by listening test (elderly and others). Experiments of automatic identification of elderly speakers showed a correct identification rate up to 95% with Gaussian Mixture Models.

More recently, Support Vector Regression with i-vectors as input has been used to estimate age from telephone speech [4], reporting a maximum absolute error (MAE) of 7.6% for both male and female speakers, with the NIST/SRE 2008/2010 database. A novel age estimation system based on Long Short-Term Memory (LSTM) recurrent neural networks (RNN) was proposed in [42]. The proposed system was tested and compared with a baseline state-of-the-art approach based on i-vectors, reporting a MAE below 8% with short duration utterances, which means a relative improvement of up to 28%.

An end-to-end recognition approach based on DNN trained on x-vectors was trained and tested on the NIST/SRE 2008/2010 corpus in [35], reporting a MAE of 4.92%. The combination of x-vectors with i-vectors resulted in MAE of 5.2%.

More recently, the Deep Learning (DL) paradigm has been applied to age estimation. For example, Deep Neural Networks (DNN) have been applied to predict both height and age of a speaker from short utterances [17]. In the case of age estimation, the Root Mean Squared Errors (RMSE) are 7.60 and 8.63 years for male and female respectively, when the mean duration of speech segments is around 2.5s. Anyway, recent studies demonstrate that age estimation is a very difficult task, even for human listeners [38], who are affected by aspects that are not related to the speaker's age, such as the listeners' own age, or the speaker's speech rate.

In [21] four approaches to estimate age and classify by gender for telephone applications are compared. Instead of estimating age using a regression analysis, people are classified into four age groups: children, youth, adults and seniors, similar to how classification will be carried out in this paper. The best automatic system, based on Hidden Markov Models (HMM), obtain good results: precision is 54%, while recall is 55%, which are not far from the results obtained by human listeners (around 60% precision and 70% recall).

The same classes were considered in [36], where a system based on DNN was trained and tested. The obtained performance was 57.53% and 88.80% for age and gender classification, respectively. The applied DNN provided the best result of speaker age recognition of existing classification methods.

Gender identification is an easier problem with state-of-the art systems achieving error rates of around 2-3% [30]. In a previous work [32], we have proposed a system based on Convolutional-Recurrent neural networks implemented using the Keras library in Python, for gender distinction and age classification into three groups: youths, adults and elderly people. A mean error below 2% was reported for gender distinction, which is comparable to the best published in the literature. A mean classification error of less than 20% was obtained for classification by age groups.

The application of DL to solve these problems relies on the use of very large networks, with a large number of parameters to adjust. Graphics Processing Units (GPUs) are used to implement DL models, to speed up training and testing. GPU memory is always limited, and the large amount of memory required by DL models can be greater than the available GPU memory [11]. Furthermore, DL models used for age estimation, or gender and emotion recognition, compete for computational resources with speech recognition and natural language processing systems in IVRs. All this encouraged us to carry out research to optimise the performance of the DL models with the smallest possible size, in order to save computational resources that could be used to solve other tasks in IVR systems.

This paper presents a framework to jointly identify speakers' gender and classify them into age groups, designed to be applied in IVR systems. The system is based on the Deep-Learning paradigm. Different network architectures and sizes are compared, to find behaviour patterns that could be helpful to determine the network size for best performance. This paper is an extended version of [33], with a more detailed analysis of the results. A preliminary study with recurrent neural networks demonstrated the usefulness of this approach [32].

The paper is structured as follows. Section 1 introduces the paper and the problem solved in this study is formulated. Section 2 presents the fundamentals of deep learning, and introduces the different network architectures analysed in this paper. The experimental work is presented in Section 3, including the speakers corpus, and the main results. The paper finishes with Section 4, where the main conclusions are presented.

2 Background on deep learning

Artificial neural networks date back to the 1940s, thanks to the work by Pitts and McCulloch [29] who studied how to simulate the way the human brain system solves general problems. The back-propagation algorithm [31] was an important milestone in this field in the 1980s, but its application brings to light very soon several problems due to training overfitting, lack of large-scale training data, and limited computational power, which made the application of artificial neural networks in real problems difficult. For decades, machine learning systems, of which neural networks are a subset, required considerable expertise to design a feature extractor that transformed raw data into a feature vector from which the learning subsystem could detect or classify patterns [19]. Deep learning emerged to overcome these problems, and its high popularity is often attributed to the following factors [43]:

- The collection of large-scale annotated training data, for image processing, speech, etc. The large size of these data sets helps to avoid overfitting when the number of network parameters is very high.
- The revolution of parallel computing, mainly due to the fast development of GPU clusters.
- Advances in the design of network structures and training algorithms. Convolutional and recurrent neural networks, temporal convolutional networks, and residual nets are some examples.

Deep learning has been used recently in many applications: video coding, fraud detection in financial applications, financial time-series forecasting, predictive and prescriptive analytics, medical image processing, and power systems research [34]. Most of the effort has been devoted to finding new applications or proposing new and more powerful architectures, but not much work has been devoted to knowing the appropriate size of a network to solve a certain problem. Deep learning has also been applied to speech analysis.

A *deep neural network* (DNN) is a feed-forward neural network with more than one hidden layer, each with a number of neurons which takes the outputs of the lower layer as an input vector, multiplies it by a weights vector and passes the result through a non-linear activation function, usually a sigmoid or hyperbolic tangent [2]. It is expressed in (1), where $\mathbf{o}^{(l-1)}$ denotes the vector of outputs of the lower layer which elements are $o_i^{(l-1)}$, $\mathbf{w}_i^{(l)}$ denotes the weights vector, $w_{0,i}^{(l)}$ is a bias term in the i-th unit, and $\sigma(x)$ is the activation function.

$$o_i^{(l)} = \sigma\Big(\sum_j o_j^{(l-1)} w_{j,i}^{(l)} + w_{0,i}^{(l)} \Big) = \sigma(\mathbf{w}_i^{(l),T} \cdot \mathbf{o}^{(l-1)}) \tag{1}$$

2.1 Convolutional neural networks

A variant of the standard neural network widely used by the image processing research community is the Convolutional Neural Network (CNN). This is a type of network that applies multiple sliding convolutional filters throughout multiple layers, and has also been used for many speech related tasks (e.g [3, 25, 28]). The CNN has a special network structure, with two kind of special layers, known as *convolution* and *pooling* layers. The input data must be organised as a number of *feature maps*, which are typically two-dimensional (2D) arrays. In the speech processing context, a 2D feature map can be obtained by applying a short-time Fourier transform (STFT), with a given number of time frames. Additional feature maps can be obtained with the same data by calculating the difference between consecutive maps, or the second difference.

A small window which elements are *known weights* slides over the 2D map. The same weights are used in each position of the window (*weight sharing*). Each feature map $\mathbf{X}_i, i = 1, \cdots, I$, is connected to a number J of feature maps $\mathbf{Y}_j, j = 1, \ldots, J$, in the convolution layer, by using a number IxJ of weight matrices $\mathbf{w}_{i,j}, i = 1, \ldots, I; j = 1, \ldots, J$. The element $y_j(m) \in \mathbf{Y}_j$ is obtained from the submatrix $\hat{\mathbf{X}}(m) \subset \mathbf{X}_i$, with the following operation:

$$\mathbf{y}_j(m) = \mathbf{w}_{i,j} \odot \hat{\mathbf{X}}(m) \tag{2}$$

where $\odot$ represents the pointwise multiplication. The number of feature maps in the convolution layer determines the number of local weight matrices to be used. The number of elements in each feature map of the convolution layer depends on how much the window matrixes overlap. To reduce the number of elements in each feature map of the convolution layer (the resolution of feature maps), a *pooling layer* is applied. A pooling function, which usually is a maximising or averaging function, is applied to several units in a local region, the size of which is determined by a parameter known as *pooling size* [2]. The scheme of a CNN layer is presented in Fig. 2.

2.2 Convolutional recurrent neural network

A Convolutional Recurrent Neural Network (CRNN) is a CNN with one or more additional recurrent layers that can find patterns in sequences. CRNNs are typically used for audio event detection/classification [6, 41]. The recurrent layer is an example of a recurrent neural network with one hidden layer. In the simplest form, it consists of a hidden state **h**, and an output **y**. At each time step t, the hidden state and the output are obtained as follows:

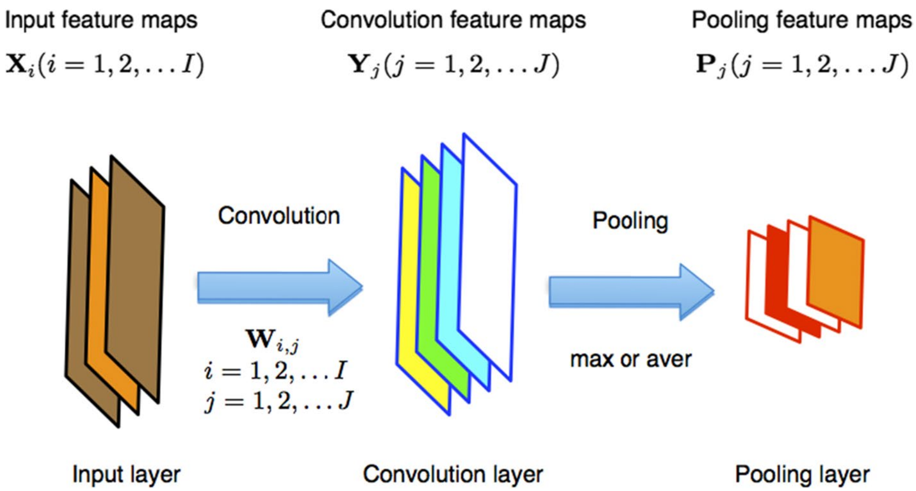

Fig. 2 Scheme of one layer in a CNN, with a convolution layer and a pooling layer (based on [2])

$$\mathbf{h}_t = f(\mathbf{W} \cdot \mathbf{h}_{t-1} + \mathbf{U} \cdot \mathbf{x}_t) \tag{3}$$

$$\mathbf{y}_t = f(\mathbf{V} \cdot \mathbf{h}_t) \tag{4}$$

where $\mathbf{W}$ is the weights matrix which feeds back the previous step hidden state, $\mathbf{U}$ is the weights matrix which connects the input to the hidden layer, and $\mathbf{V}$ is the weights matrix which connects the hidden layer to the output. The recurrent neural network can be trained using the back-propagation algorithm. The vanishing gradient problem arises due to the structure, and the network can only remember the latest information and not the previous one [16].

To address the vanishing gradient problem, the *Long short term memory* (LSTM) was proposed in [13]. It is made up of a cell, an input gate, an output gate and a forget gate. The cell is used to store information from the beginning to the end of training, to avoid the vanishing gradient problem. The other gates are used to regulate the flow of information in and out of the cell.

The Gated Recurrent Unit (GRU) [7], is based on the LSTM, but uses fewer parameters, making the implementation easier. The following equation describes how the GRU works. First, $\mathbf{r}_t$ is computed:

$$\mathbf{r}_t = \sigma(\mathbf{W}_r \cdot \mathbf{h}_{t-1} + \mathbf{U}_r \cdot \mathbf{x}_t) \tag{5}$$

being, $\mathbf{W}_r, \mathbf{U}_r$ weight matrices, and σ the activation function (logistic sigmoid function). The new remember $\hat{\mathbf{h}}_t$ is computed from $\mathbf{r}_t$ with (6):

$$\hat{\mathbf{h}}_t = tanh(\mathbf{W} \cdot (\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{U} \cdot \mathbf{x}_t) \tag{6}$$

The hidden state value is finally updated using $\hat{\mathbf{h}}_t$, the previous hidden state, and the value $\mathbf{z}_t$, calculated as follows:

$$\mathbf{z}_t = \sigma(\mathbf{W}_z \cdot \mathbf{h}_{t-1} + \mathbf{U}_z \cdot \mathbf{x}_t) \tag{7}$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \hat{\mathbf{h}}_t \tag{8}$$

These recurrent neural networks are used in speech processing, because they can find patterns in sequences. They are used to process very long non-stationary signals, such as speech and audio signals.

2.3 Temporal convolutional networks

The main disadvantage of the approach based on CNN and RNN (LSTM or GRU) is that it requires two separate models to process time series. *Temporal Convolutional Networks* (TCN) are a new type of network which can also be used to model sequences, but using only convolutional layers. They have been proposed in the seminar paper by Lea et al. [18], and have been used recently for speech enhancement [26].

The architecture of a TCN is like that of a CNN, but performing dilated convolutions in 1D with causal filters followed by max-pooling layers. The main characteristics of TCN are the following:

- The convolutions in the architecture are causal. The TCN uses causal convolutions, where an output at time t is obtained with elements from time t and earlier in the previous layer.
- The model can map a sequence of any length to an output of the same length. This is also possible with RNN. The TCN uses a 1D fully convolutional network architecture, where the hidden layers and the input layer are the same length. This is achieved with 1D convolutions by using zero padding of length equal to the kernel filter size minus one, to keep subsequent layers the same length as previous ones.
- Simple causal convolutions only look back at history with size proportional to the depth of the network. To avoid this, the TCN employs dilated convolutions that allow an exponentially large receptive field.

In this paper, we have combined CNN and TCN, and compared architectures built with CNN + RNN, with architectures built with CNN + TNN, and different sizes, to obtain information on the best architecture to use and the size above which there is no performance improvement.

3 Experimental work

3.1 Speech corpus

The Mozilla Common Voice dataset has been used, an open and publicly available multi-language speech corpus. The English corpus used in this work is composed of 12GB of labeled audio files, containing thousands of speech clips (most are short phrases between 2-5s), recorded and annotated by anonymous speakers. Labels include speakers' ages grouped by decades (from 'teens' to 'octogenarians') and gender. This labelling prevents the implementation of an age estimation system based on regression networks. In total, there are 143,170 audio files (≈ 98h) with age and gender labels. The age distribution is very unbalanced, as can be seen in Fig. 3, where only 0.3% of the total audio files (432 files) correspond to octogenarians.

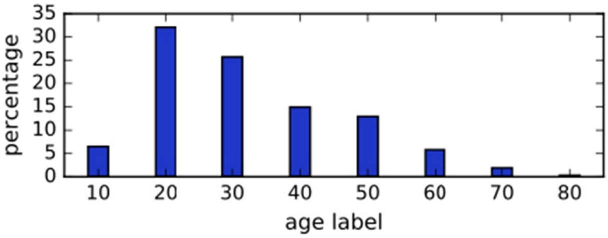

Fig. 3 Dataset age group distribution

For IVR applications, where the main objective is to route callers to specialised agents, classification into age groups of decades is not desired, because call centres do not train agents so specifically. Instead, it is reasonable to train agents to attend to young people, adults, and seniors. Because of that, the audio files have been grouped into three age groups: young people (teens and twenties), adults (thirties, forties and fifties), and seniors (sexagenarians, heptagenarians and octogenarians). The system is built to have four outputs, one for each age group and a final one for gender classification. Other works consider six groups, corresponding to the combination of age group and gender.

The original audio files are in MP3 format, with a bit rate of 64kbps. They are transcoded to PCM with a sampling rate of 8kHz to approximate the quality of telephone lines. Losses due to MP3 coding mainly affect the high frequency bands of audio signals, and most of the information below 4kHz remains available. The audio quality throughout the corpus is highly variable. Some files are almost unintelligible (due to noise, saturation, etc.) while other files are of very good quality. We have considered that including all the files makes the evaluation more realistic for real life applications, such as IVRs, where the quality of recordings is sometimes very low.

It is worth commenting that some audio files contain long silences at the beginning and the end, which have been detected using a voice activity detector and deleted to avoid confusion. Additionally, all the audio signals have been normalised, so that their amplitudes are in the range ± 1.

Audio files are divided into one-second blocks (with 50% overlapping between adjacent blocks). The Short-time Fourier transform (STFT) is applied to each block, with 20ms length frames (160 samples), windowed with the hamming window, and with adjacent windows overlapping 50% of their length. After STFT calculation, the energy of 20 subbands in the Mel-scale is obtained for each frame. Zero padding is used to complete the last block of each audio file. Therefore, each second of the audio file is represented by a [100x20] matrix. Data are normalised to have zero-mean and unit-variance according to the mean and standard deviation of the training set.

3.2 Experimental settings

The dataset was divided into three subsets for the experiments: training set, which contains 70% of the available files; validation set, with 15% of the files, and test set, with the remaining 15% of the files. The size of these sets was empirically determined, to have a large number of data for training, and enough data for testing. Due to privacy concerns, speakers' ID labels are not included in the corpus. For this reason, it is not possible to ensure that a particular speaker is not represented in the different sets, but it is unlikely.

Table 1 shows the different networks tested in this work, detailing the layer structure of each network. The number of units in each layer is indicated by the number preceding the layer type. Additionally dropout regularisation with a dropout rate of 0.2 is used between every layer for all the networks.

We opted to fix the input length to simplify the classification. However, since the system needs to work with varying signal lengths, a simple time-averaging strategy is used. This way, audio signals that exceed the selected input length are divided into multiple segments to be processed independently by the network, and their outputs are later combined by averaging, or with the maximum voting strategy.

The different network types are made up of a combination of the following layers:

- *Conv*: 2D Convolutional layer with rectified linear unit (ReLU) activation function. A 2D convolutional layer applies sliding convolutional filters to the input. The layer convolves the input by moving the filters along the input vertically and horizontally and computing the dot product of the weights and the input, and then adding a bias term. The size of the kernel (filter) is indicated by the numbers in brackets. This layer is followed by max-pooling in the feature axis with a poolsize of 2. Max-pooling is a sampling-based discretisation process to reduce the dimensionality of the 2D Convolutional layer.
- *R*: Reshaping layer, used to change the dimensions of its input, without changing its data. It turns the 4 dimensional tensor ([batch x time x features x channels]) into a 3 dimensional tensor by combining the feature and channel dimensions into a single one.
- *GRU*: Gated Recurrent Unit [7], with *tanh* activation and sigmoid recurrent activation. The GRU aims to solve the vanishing gradient problem that appears in a standard recurrent neural network. It returns the last output in the output sequence.

Table 1 Tested neural networks detailing their architecture

#	Type	Network Architecture
1	CRNN	48 Conv(3,3)+24 Conv(3,3)+R+48 GRU+O
2	CRNN	48 Conv(3,3)+24 Conv(3,3)+R+64 GRU+O
3	CRNN	48 Conv(3,3)+24 Conv(3,3)+R+96 GRU+O
4	CRNN	48 Conv(3,3)+24 Conv(3,3)+R+112 GRU+O
5	CRNN	48 Conv(3,3)+24 Conv(3,3)+48 Conv(3,5)+R+48 GRU+O
6	CRNN	48 Conv(3,3)+24 Conv(3,3)+64 Conv(3,5)+R+96 GRU+O
7	CTCN	48 Conv(3,3)+24 Conv(3,3)+32 Conv(3,5)+R+32 TCN+DC1D (4)+MP+O
8	CTCN	48 Conv(3,3)+24 Conv(3,3)+48 Conv(3,5)+R+48 TCN+DC1D (4)+MP+O
9	CTCN	48 Conv(3,3)+24 Conv(3,3)+48 Conv(3,5)+R+56 TCN+DC1D (4)+MP+O
10	CTCN	48 Conv(3,3)+24 Conv(3,3)+R+32 TCN(5x4)+DC1D(4)+MP+O
11	CTCN	48 Conv(3,3)+24 Conv(3,3)+R+54 TCN(5x4)+DC1D(4)+MP+O
12	TCN	32 TCN+DC1D(4)+MP+O
13	TCN	48 TCN+DC1D(4)+MP+O
14	TCN	64 TCN+DC1D(4)+MP+O
15	CNN	48 Conv(3,3)+24 Conv(3,3)+R+DC1D (2)+MP+O
16	CNN	48 Conv(3,3)+24 Conv(3,3)+48 Conv(3,5)+R+DC1D(4)+MP+O
17	CNN	48 Conv(3,3)+24 Conv(3,3)+96 Conv(3,5)+R+DC1D(3)+MP+O
18	CNN	48 Conv(3,3)+32 Conv(3,3)+128 Conv(3,5)+R+DC1D(3)+MP+O

- *TCN*: Temporal Convolutional Network made up of 4 increasingly dilated 1D convolutional layers (dilation $\in \{1, 2, 4, 8\}$) with residual connections. 1D convolutions have a kernel size of 5 with ReLu activation. When the input size is different to the number of filters, a residual block is used (the output of a layer is taken and added to another layer deeper in the block, with linear activation).
- *DC1D*: Depthwise Convolution 1D, a type of convolutional layer which acts on each input channel separately. ReLU activation, kernel size fixed to 5, and depth multiplier (number of filters per channel) are indicated in brackets.
- *MP*: Maximum pooling performed only in the time axis, with poolsize equal to the sequence length.
- *O*: Output layer: 4 units fully connected with the previous layer, with sigmoid activation..

All the networks were trained for 300 epochs using ADAM optimisation with 0.001 learning rate and binary cross-entropy as loss function. Mini batch size was fixed to 2048. To counteract class imbalance, errors due to the different classes in the training set are weighted according to the size of the age group (0.25 for young people, 0.2 for adults and 0.55 for seniors). The best network parameters were selected in terms of the minimum validation loss. Neural networks were implemented in Python 3.5 using Keras 2.24 [8] with Tensorflow r1.12 [1] as the backend.

3.3 Results and discussion

The models have been compared in terms of block-level performance and file-level performance. Block-level results are obtained by evaluating one-second audio blocks independently, while file-level results are obtained by averaging the outputs of all blocks that make up each file before computing the metrics.

Table 2 shows the error for gender classification of the tested networks both at block and file level. The files available in the dataset have been divided into three subsets, one for training, another for validation, and a third for testing. The error rate is calculated with the test set, considering the correct age group of the speaker and his gender. An error is computed when the speaker is classified in a different age group or gender. The 95% confidence intervals of the error rates estimated using the Clopper-Pearson method have also been included. Although the results are always very good, comparable to the best results published in the literature with the same or similar corpus, working at file-level seems to be better, reducing the average error by approximately 35% on average. The cost is paid by processing more segments of the audio file, in order to obtain values that are averaged.

The results are better for larger networks, but not better enough to justify the higher computational cost. Additionally, there is not a clear relationship between the type of architecture and the results. Similar results are obtained with CNN, TCN, CTCN, and CRNN, although theoretically, architectures that include TCN and RNN are more appropriate for processing speech signals. The main conclusion is that feed-forward networks are enough for this application, but the larger the network size, the better the results.

To compare the results with those obtained with standard methods described in the literature, two different techniques have been implemented and applied to the same database. In both cases the results presented in this paper outperforms those obtained by these two methods described in the literature:

Table 2 Error rate and 95% confidence intervals for gender classification with the tested neural networks as a function of the number of parameters of the network

#	Params	Block-level Error	File-level Error
1	35884	3.22% [3.11% - 3.33%]	2.06% [1.87% - 2.26%]
2	47132	3.13% [3.02% - 3.24%]	2.07% [1.89% - 2.27%]
3	74236	2.08% [1.99% - 2.17%]	1.87% [1.69% - 2.06%]
4	90092	2.74% [2.64% - 2.84%]	1.76% [1.59% - 1.94%]
5	59260	3.21% [3.10% - 3.32%]	2.15% [1.96% - 2.35%]
6	94156	2.92% [2.81% - 3.03%]	2.03% [1.85% - 2.23%]
7	44828	3.43% [3.32% - 3.55%]	2.17% [1.98% - 2.37%]
8	77164	3.15% [3.04% - 3.26%]	1.98% [1.80% - 2.17%]
9	94756	2.99% [2.88% - 3.10%]	1.82% [1.65% - 2.01%]
10	51580	3.22% [3.11% - 3.33%]	2.00% [1.82% - 2.20%]
11	97054	2.95% [2.84% - 3.06%]	1.79% [1.62% - 1.97%]
12	21028	4.13% [4.01% - 4.26%]	2.66% [2.45% - 2.88%]
13	42484	3.64% [3.52% - 3.76%]	2.25% [2.05% - 2.46%]
14	84484	3.17% [3.06% - 3.28%]	1.86% [1.68% - 2.05%]
15	14236	4.25% [4.12% - 4.38%]	2.81% [2.59% - 3.04%]
16	30700	3.61% [3.49% - 3.73%]	2.27% [2.08% - 2.48%]
17	49372	3.40% [3.29% - 3.52%]	2.09% [1.90% - 2.29%]
18	81028	3.13% [3.02% - 3.24%]	1.89% [1.71% - 2.08%]
SVR (baseline system [17])		12.88% [12.67% - 13.09%]	8.92% [8.54% - 9.31%]
Random Forest [37]		6.95% [6.79% - 7.11%]	5.08% [4.79% - 5.38%]

- In a first approach, we compare with the baseline system described in [17], which consists in applying Support Vector Regression with i-vectors obtained from a 256 Gaussian Mixture Model applied to 20 MFCCs, Δ-MFCCs, and Δ^2-MFCCs. With this method we obtained 12.88% of error rate working at block-level, and 8.92% of error rate working at file-level.
- On the other hand, the method described in [37], which consists in using a random forest generated using 100 trees with the features vectors generated from a 1024 Gaussian Mixture Model applied to 64 MFCCs obtained from a time window of 25ms. With this methodology we obtain a 6.95% of error rate working at block-level and 5.08% of error rate working at file-level over the same database.

Table 3 shows the age group classification error of the tested networks at both block and file levels. Again, the 95% confidence intervals of the error rates estimated using the Clopper-Pearson method have also been included. The file-level error is significantly lower. In relative terms, age group classification error is reduced by 18% on average. Focusing on the results at file-level, it is interesting to study the performance of the different kinds of networks, and the dependence on the number of parameters. Figure 4 depicts the file-level age classification error, as a function of the number of parameters in the network.

Results show a clear relation between the number of free parameters to be adjusted during training and the performance of the different networks. Again, the larger the network size, the better the results. The type of network is also important, and the

Table 3 Mean classification error and 95% confidence intervals with the tested neural networks described in Fig. 1 classifying in age groups, as a function of the number of parameters of the network

#	Params	Block-level Error	File-level Error
1	35884	27.52% [27.24% - 27.80%]	23.20% [22.63% - 23.77%]
2	47132	25.22% [24.95% - 25.49%]	21.14% [20.60% - 21.69%]
3	74236	22.68% [22.42% - 22.94%]	18.13% [17.62% - 18.65%]
4	90092	21.72% [21.46% - 21.98%]	16.77% [16.27% - 17.28%]
5	59260	26.23% [25.95% - 26.51%]	21.92% [21.37% - 22.48%]
6	94156	24.04% [23.77% - 24.31%]	19.64% [19.11% - 20.18%]
7	44828	27.98% [27.70% - 28.26%]	23.21% [22.65% - 23.78%]
8	77164	25.53% [25.26% - 25.81%]	19.93% [19.40% - 20.47%]
9	94756	24.47% [24.20% - 24.74%]	19.72% [19.19% - 20.26%]
10	51580	26.78% [26.50% - 27.06%]	21.23% [20.68% - 21.78%]
11	97054	22.62% [22.36% - 22.88%]	17.04% [16.54% - 17.55%]
12	21028	37.30% [37.00% - 37.61%]	31.37% [30.75% - 32.00%]
13	42484	31.35% [31.06% - 31.64%]	28.76% [28.15% - 29.37%]
14	84484	25.79% [25.52% - 26.07%]	20.12% [19.59% - 20.66%]
15	14236	36.16% [35.86% - 36.46%]	32.01% [31.38% - 32.64%]
16	30700	30.14% [29.85% - 30.43%]	24.90% [24.32% - 25.49%]
17	49372	27.65% [27.37% - 27.93%]	22.34% [21.79% - 22.90%]
18	81028	25.53% [25.26% - 25.81%]	19.57% [19.04% - 20.11%]
SVR (baseline system [17])		44.73 % [44.42% - 45.04%]	44.56% [43.90% - 45.23%]
Random Forest [37]		40.29% [39.98% - 40.60%]	37.57% [36.92% - 38.22%]

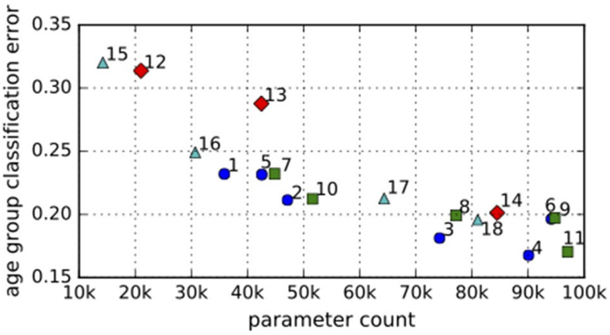

Fig. 4 Age group classification error at file-level as a function of network size

best results are obtained with the combination of CNN and RNN or TCN. The results obtained by the CRNN and CTCN networks show that, in this particular application, the use of 2 rather than 3 convolutional layers preceding the temporal memory layer is preferred, most likely due to the higher dimension of the reshaped data.

Again, we also want to compare the obtained results with those generated by standard methods described in the literature. Using the baseline method described in [17] we obtain an error of 44.73% working at block-level, and an error of 44.56% working at file-level, while using the method described in [37] we obtain an error rate of 40.29%

Table 4 Six classes results for best performing neural network of each type (Recall / Precision / F1 score)

#	Type	Params	Block-level			File-level		
			Recall	Precision	F1	Recall	Precision	F1
4	CRNN	90092	0.76	0.81	0.78	0.74	0.81	0.77
11	CTCN	97054	0.72	0.79	0.76	0.74	0.83	0.78
14	TCN	84484	0.66	0.72	0.69	0.73	0.82	0.77
18	CNN	81028	0.69	0.76	0.73	0.70	0.78	0.74

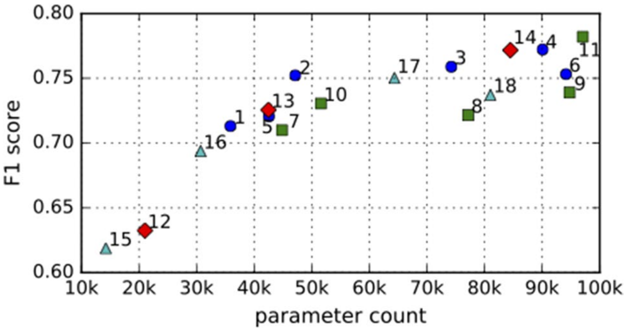

Fig. 5 F1 score at file-level as a function of network size

working at block-level and 37.57% working at file-level. Again, the proposed deep learning based methodologies outperform the results obtained with classical methods.

We have also built, trained and tested neural networks with six outputs, representing the following six categories: Young Male (YM), Young Female (YF), Adult Male (AM), Adult Female (AF), Senior Male (SF) and Senior Female (SF). Table 4 shows the results of the best performing network of each type for the 6-class problem in terms of Recall, Precision and F1 score, and Fig. 5 depicts the F1 score at file-level for all networks tested. These results indicate that when the system is implemented in a real IVR system, around 80 % of the callers will be routed to the correct specialised agent.

These results, compared to the age group classification error presented in Table 2, show an increase in the average error of about 2% for block-level and 1.3% for file-level.

Greater differences are observed among the different types of networks, with worse results for the CNN and better results for the TCN or CRNN. TCN is the type of network that obtains the greatest improvement working at the file-level compared to the block-level (Table 4). These results indicate that neural networks with structures that allow finding patterns in temporal sequences are better for processing speech signals in this application. A very interesting result is the better performance of TCN working at the file-level, rather than at the block-level, which indicates that this type of neural network performs comparatively better with long signals than with short signals.

The results presented as mean error in the classification by age group or by gender allow us to draw conclusions about the general performance of each system. But we must not forget that the age groups are unbalanced and that this could have some influence, causing the errors to be concentrated in the worst represented group. To avoid this, the error was weighted so that the one made by the worst represented group had more importance. To know if this strategy has been successful, it has been considered pertinent to present a confusion matrix of the different age and gender groups.

Figure 6 shows the relative confusion matrix for the overall best performing network (network number 4). The confusion matrix shows that most of the errors come from confusion between adjacent age groups, such as young and adults, or adults and seniors, but the error is low between young and senior people groups. Gender misclassification errors are much more unlikely. One of the most interesting results shown in the confusion matrix is that there are no great differences among the percentage of error for the different age groups, despite the fact that seniors are worse represented in the training set.

4 Conclusions

This paper presents a study of age and gender recognition from speech using different types of neural networks. The underlying application is IVR systems, where gender and age classification can be used for demographic analysis and to optimise call flow, for example by offering immediate assistance from a human operator or trying to react with appropriate dialogue strategies. This is an example of data-directed routing, being data obtained directly from the speech utterance during the call.

One of the main difficulties for implementing this kind of systems is access to a dataset that could be used to train and test the classifiers. For this study one of the few freely available datasets with information about speakers' gender and age has been used, the Mozilla Common Voice dataset, an open-access corpus that anyone can use to train speech-enabled applications. This dataset doesn't contain the exact speaker age, but information about the decade the speaker belongs to. Therefore, instead of age prediction, the classification of speakers into age groups is carried out. This is compatible with real IVR systems, where agents can be trained to serve callers based on age group.

Each audio file was pre-processed to obtain the matrix which feeds the classifier. Audio files were divided into one-second blocks (with 50% overlapping between adjacent blocks),

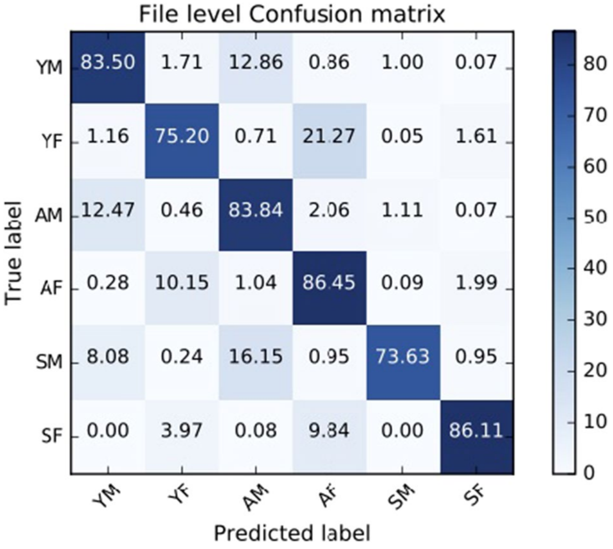

Fig. 6 Relative confusion matrix (rows sum up to 100%) at file-level for network number 4

and each block was segmented into 20ms windows (160 samples), with adjacent windows overlapping 50% of its length. Spectral features were obtained for each segment, and the information to be applied to the classifier for each block is a [100x20] matrix, being 100 the number of segments, and 20 the number of features of each segment.

Different kinds of neural networks have been tested. All of them can be included in the Deep Neural Networks group. We have taken into account Convolutional Neural Networks, which are a kind of feed-forward neural network, and networks proposed to process time series considering historical information, such as Recurrent Convolutional Neural Networks and Temporal Convolutional Networks. The results confirm that all types of networks obtain good results for gender classification, but the combination of Convolutional Networks and Temporal Neural Networks performs better for age group classification. The influence of the number of free parameters on the classifier performance has also been studied, demonstrating that the larger the network size (number of free parameters), the better the performance in both applications. However, it is not worth using more than 50 thousand parameters, regardless of the type of network configuration used (the behaviour of different networks with an equivalent number of free parameters is similar). An important topic that will be studied in future works is the importance of data being unbalanced.

The results for gender classification are really good, with errors commonly below 2%, which is comparable to the best published in the literature. The results for age group classification are also promising, with the best configurations achieving a classification error below 20%. In terms of precision and recall, the best performing systems achieve around 80% and 70% respectively, which is in line with human performance according to the literature. These results indicate that gender and age groups classification can be implemented with very good results to improve the service provided in IVR systems.

Despite the fact that the results obtained in terms of mean error were very good, it remained to be known whether the error in the estimation of the age group was distributed more or less uniformly among the different groups, despite the fact that the group of seniors is very poorly represented in the database. An analysis of the confusion matrix allows us to verify that there are no great differences in the mean error in each age group. This has been achieved thanks to the fact that the error function for training gives more importance to the errors made in the worst represented groups.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

References

1. Abadi M, Agarwal A, Barham P, et al (2015) TensorFlow: large-scale machine learning on heterogeneous systems. http://tensorflow.org/. Software available from tensorflow.org

2. Abdel-Hamid O, Abdel-Rahman M, Jiang H, Deng L, Penn G, Yu D (2014) Convolutional neural network for speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing 22(10):1533–1545
3. Badshah A, Ahmad J, Rahim N, Baik S (2017) Speech emotion recognition from spectrograms with deep convolutional neural network. In: 2017 International conference on platform technology and service (PlatCon), pp 1–5
4. Bahari M, McLaren M, Van Leeuwen D, et al (2012) Age estimation from telephone speech using i-vectors. In: Proceedings of Interspeech 2012. Portland, USA
5. Bhat C, Mithum B, Saxena V, Kulkarni V, Kopparapu S (2013) Deploying usable speech enabled ivr systems for mass use. In: 2013 IEEE international conference on human computer interaction (ICHCI), pp 1–5
6. Cakir E, Adavanne S, Parascandolo G, Drossos K, Virtanen T (2017) Convolutional recurrent neural networks for bird audio detection. In: 2017 25th European signal processing conference (EUSIPCO), pp 1744–1748
7. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv:1406.1078
8. Chollet F, et al (2015) Keras. https://keras.io
9. Couper M, Singer E, Tourangeau R (2004) Does voice matter? An interactive voice response (IVR) experiment. Journal of Official Statistics 20(3):551–570
10. Devillers L, Vidrascu L (2006) Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In: INTERSPEECH 2006. International Speech Communication Association, pp 801–804.
11. Gao Y, Liu Y, Zhang H, Li Z, Zhu Y, Lin H, Yang M (2020) Estimating GPU memory consumption of deep learning models. In: Proceedings of the 28th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering. ACM, pp 1342–1352
12. Gorin A, Riccardi G, Wright J (1997) How may I help you? Speech Communication 23:113–127
13. Hochreiter S, Schmidhuber J (1997) Long short term memory. Neural Computation 9(8):1735–1780
14. Huang J, Li B, Zhu J, Chen J (2017) Age classification with deep learning face representation. Multimedia Tools and Applications 76:20231–20247
15. Ilyas M, Othmani A, Nait-ali A (2020) Auditory perception based system for age classification and estimation using dynamic frequency sound. Multimedia Tools and Applications 79:21603–31626
16. Jinglong C, Hongjie J, Yanhong C, Qian L (2019) Gated recurrent unit based recurrent neural network for remaining useful life prediction of nonlinear deterioration process. Reliability Engineering and System Safety 185:372–382
17. Kalluri SB, Vijayasenan D, Ganapathy S (2019) A deep neural network based end to end model for joint height and age estimation from short duration speech. In: 2019 IEEE International conference on acoustics, speech and signal processing (ICASSP 2007). IEEE, pp 6580–6584
18. Lea C, Vidal R, Reiter A, Hager GD (2016) Temporal convolutional networks: a unified approach to action segmentation. In: European conference on computer vision. Springer, pp 47–54
19. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 323:436–444
20. Mehrbod N, Grilo A, Zutshi A (2018) Caller-agent pairing in call centers using machine learning techniques with imbalanced data. In: 2018 IEEE International conference on engineering, technology and innovation (ICE/ITMC). IEEE, pp 1–6
21. Metze F, Ajmera J, Englert R, Bub U, et al (2007) Comparison of four approaches to age and gender recognition for telephone applications. In: 2007 IEEE International conference on acoustics, speech and signal processing (ICASSP 2007), vol 4, pp IV–1089
22. Minematsu N, Sekiguchi M, Hirose K (2002) Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers. In: 2002 IEEE International conference on acoustics, speech, and signal processing (ICASSP 2002), vol 1, pp I–137
23. Mohino-Herranz I, García-Gómez J, Utrilla-Manso M, Rosa-Zurera M (2018) Precision maximization in anger detection in interactive voice response systems. In: 145th convention of the audio engineering society, paper number, pp 10090
24. Mubarak E, Shahid T, Mustafa M (2020) Does gender and accent of voice matter?: an interactive voice response (ivr) experiment. In: Proceedings of the 2020 international conference on information and communication technologies and development. ACM Digital Library, pp 739–746
25. Neumann M, Vu NT (2017) Attentive convolutional neural network based speech emotion recognition: a study on the impact of input features, signal length, and acted speech. arXiv preprint arXiv:1706.00612

26. Pandey A, Wang D (2019) Tcnn: temporal convolutional neural network for real-time speech enhancement in the time domain. In: 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP 2019), pp 6875–6879
27. Pappas D, Androutsopoulos I, Papageorgiou H (2015) Anger detection in call center dialogues. In: 2015 6th IEEE international conference on cognitive infocommunications (CogInfoCom), pp 139–144
28. Park SR, Lee JW (2017) A fully convolutional neural network for speech enhancement. In: Proc. Interspeech, pp 1993–1997
29. Pitts W, McCulloch W (1947) How we know universals the perception of auditory and visual forms. Bull Math Biophys 9(3):127–147
30. Ranjan S, Hansen JH (2017) Improved gender independent speaker recognition using convolutional neural network based bottleneck features. In: Proceedings of Interspeech, pp 1009–1013
31. Learning representations by back-propagating errors (1986) Rumelhart, D., al. Nature 521:533–536
32. Sánchez-Hevia H, Gil-Pita R, Utrilla-Manso M, Rosa-Zurera M (2019) Convolutional-recurrent neural network for age an gender prediction from speech. In: 2019 signal processing symposium, krakow (Poland). IEEE, pp 246–249
33. Sánchez-Hevia H, Gil-Pita R, Utrilla-Manso M, Rosa-Zurera M (2020) Age and gender recognition from speech using deep neural networks. In: Advances in Physical Agents II. Proceedings of the 21st International Workshop of Physical Agents (WAF 2020). Advances in Intelligent Systems and Computing Series. Springer Nature Switzerland, pp 332–344
34. Sengupta S, Basak S, Saikia P, Sayak P, Tsalavoutis V, Atiah F, Ravi V, Peters A (2020) A review of deep learning with special emphasis on architectures, applications and recent trends. Knowledge-Based Systems 194(105596):1–33
35. Ghahremani P, Nidadavolu PN, Chen N, Villalba J, Povey D, Khudanpur S, Dehak N (2018) End-to-end deep neural network age estimation. In: Proceedings of the 19th annual conference of the international speech communication association, INTERSPEECH 2018. ISCA, pp 277–281
36. Markitantov M, Verkholyak O (2019) Automatic recognition of speaker age and gender based on deep neural networks. In: Speech and computer, LNAI, vol 11658. Springer Nature, pp 327–336
37. Singh R, Raj B, Baker J (2016) Short-term analysis for estimating physical parameters of speakers. In: 2016 4th international conference on biometrics and forensics (IWBF). IEEE, pp 1–6
38. Tsang K, Wong K, Kang Y (2020) Age estimation in short speech utterances based on lstm recurrent neural networks. Toronto Working Papers in Linguistics(TWPL) 42:1–10
39. Vidrascu L, Devillers L (2006) Real-life emotion representation and detection in call centers data. In: International conference on affective computing and intelligent interaction. Springer, pp 739–746
40. Wang M, Wang X (2010) Study on the workforce scheduling and routing strategies of heterogeneous agents in call centers. In: Advances in economics, business and management research, vol 159 (Fifth International Conference on Economic and Business Management). Atlantic Press, pp 577–583
41. Xu Y, Kong Q, Wang W, Plumbley M (2018) Large-scale weakly supervised audio classification using gated convolutional neural network. In: 2018 IEEE International conference on acoustics, speech and signal processing (ICASSP 2018), pp 121–125
42. Zazo R, Nidadavolu P, Chen N, Gonzalez-Rodriguez J, Dehak N (2018) Age estimation in short speech utterances based on lstm recurrent neural networks. IEEE Access 6:22524–22530
43. Zhao ZQ, Zheng P, Xu ST, Wu X (2019) Object detection with deep learning: a review. IEEE Transactios on Neural Networks and Learning Systems 30(11):3212–3232