Cross-Cultural Emotion Recognition and Comparison Using Convolutional Neural Networks

Alexander Iliev[1,2], Ameya Mote[1], Arjun Manoharan[1]

[1] SRH University Berlin, Charlottenburg, Germany,
[2] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria
ailiev@berkeley.edu, 3104966@stud.srh-campus-berlin.de,
3105614@stud.srh-campus-berlin.de

Abstract. The paper sets to define a comparison of emotions across 3 different cultures namely Canadian French, Italian, and North American. This was achieved using speech samples for each of the three languages subject to our study. The features used were MFCCs and were passed through convolutional neural network in order to verify their significance for the task of emotion recognition through speech. Three different systems were trained and tested, one for each language. The accuracy came to 71.10%, 79.07%, and 73.89% for each of them respectively. The aim was to prove that the feature vectors we used were representing each emotion well. A comparison across each emotion, gender and language was drawn at the end and it was observed that apart from the emotion *neutral*, every other emotion was expressed somewhat differently by each culture. Speech is one of the main vehicles to recognize emotions and is an attractive area to be studied with application to presenting and preserving different cultural and scientific heritage.

Keywords: Emotion Recognition, Speech Analysis, Language Processing, Convolutional Neural Networks, Cultural Comparison.

1 Introduction

In our lives we often face difficulties when trying to understand the emotions expressed by others. This could be much harder when an emotional expression by different cultures and languages is added to the picture. Hence, this scenario can often produce misunderstandings and can lead to miscommunications. It follows therefore, that if a system that involves emotion recognition (Kwonl, Chan, Hao, & Lee, 2003) is put in place, the emotion recognition across different cultures can be better understood. Every emotion can be perceived easily but to apply them in real world applications becomes a daunting task. The purpose of this project was therefore to narrow the gap between emotions expressed in different cultures and compare each individual one. This result could further help explain how closely related people are with each other and can automate services and create different technology for the presentation and preservation of cultural and scientific heritage.

Digital Presentation and Preservation of Cultural and Scientific Heritage. Conference Proceedings. Vol. 10, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS, 2020. ISSN: 1314-4006, eISSN: 2535-0366

2 Problem Description

A case study carried out by Alison Owens of Central Queensland University (Owens, 2008), defined cross cultural communication issues between Filipino staff and their Australian customers. The study, over a period of phone calls for months, found out how cultural difference affected the performance of staff and the satisfaction rating of the customers. The results showed that if there was a better understanding of the opposite culture through cross-cultural training, staff performance would have been better, ultimately leading to happier customers. Thus, this paper defined a way of comparing the emotional differences and helped understanding individual emotions better.

Speech helps in conveying important and relevant information to the listeners so that they can recognize the intentions of the speaker in front of them. Emotions play an equally important role in verbal communication (Iliev & Stanchev, Smart Function Digital Content Ecosystem using Emotion Analysis of Voice, 2017). Different emotions could be used to keep the audience interested and attentive.

There are a lot of cultures in the world, so even selecting three distinct cultures proved to be problematic since, there is a need to have common ground for the basis of comparison. After much deliberation, Canadian French (Livingstone & Russo, 2018), Italian (Costantini, Iaderola, Paoloni, & Todisco, 2014), and North American culture (Livingstone & Russo, 2018) datasets were chosen as they offered a good diversity. Each dataset was comprised of emotions, acted by professional actors. Since a combined dataset was not available, the three different datasets were merged to make one single dataset, which suited our requirements.

Annotation proved another big obstacle, as each dataset comprised of several files. Since each dataset was collected through different means, they needed to be labelled properly so that they could be used for further research as well.

3 Emotion Recognition

In the 1970's the psychologist Paul Eckman defined the six basic emotions as:

Table 1. Basic Emotion Types

Basic Emotions
1. Happiness
2. Anger
3. Sadness
4. Disgust
5. Fear
6. Surprise

Another, 7th emotion was added later - *neutral*. There is different literature that suggest various number of emotions. One suggested that there are 27 different emotions that were obtained by combining the traditional emotions (Cowen & Keltner, Sept, 05, 2017). Other research uses three emotions just fine (Iliev & Scordilis, Emotion Recognition in Speech using Inter-Sentence Glottal Statistics, 2008), but this usually depends on the type of task. In our project we used four primary emotions (Iliev & Scordilis, Spoken Emotion Recognition Using Glottal Symmetry, 2011) and the cross-cultural comparison was based on them as explained below. Naturally, the way of expressing a single emotion by individuals across different cultures varies greatly and thus, it is harder to understand them at times. Usually a given dataset can comprise of audio, visual, or audio-visual information. Emotion recognition has improved in recent years and many big corporations like Amazon, Facebook and others are using sentiment and emotion analysis techniques to understand their customers better and make decisions based on the results.

4 System Architecture

For the best classification performance, it is necessary to have a similar representation of features for any speech. Their types can vary significantly (Iliev & Stanchev, Glottal Attributes Extracted from Speech with Application to Emotion Driven Smart Systems, 2018). Mel-Frequency cepstral coefficient helps greatly in achieving this. It is so because the human ear cannot perceive audio signals linearly above 1000 Hz. For this, Mel scale is provided with 2 types of filters: below 1 kHz, they are spaced linearly whereas above 1 kHz, there are spaced logarithmically from one another (Iliev A. I., 2012). Thus, Mel scale is a nonlinear representation of the signals in human ear which are above 1kHz. Mel frequency is obtained by:

$$F = [1000/\log_{10}(2)] * [\log_{10}(1 + \frac{f}{1000})] \tag{1}$$

where F stands for non-linear Mel-Frequency and f stands for frequency is Hz. The formula and the diagram given below depict how we can obtain the MFCCs:

$$MFCC_m = \sum_{k=1}^{13}(\log_{10} \mathrm{E_K})[\frac{m\left(k-\frac{1}{2}\right)\pi}{20}] \tag{2}$$

Where m =1, 2, … N, N is the total number of Mel cepstral coefficients and E_k stands for the energy output for the k^{th} filter, where k = 1, 2, … 13 (Iliev A. I., 2012).

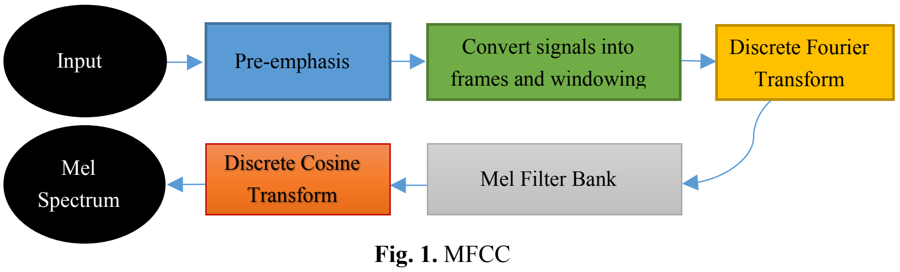
Input
Pre-emphasis
Convert signals into frames and windowing
Discrete Fourier Transform
Mel Filter Bank
Discrete Cosine Transform
Mel Spectrum

Fig. 1. MFCC

4.1 Databases and Pre-processing

As explained previously, the dataset used from features were extracted are of Canadian French (Gournay, Lahaie, & Lefebvre, 2018), Italian (Costantini, Iaderola, Paoloni, & Todisco, 2014), and North American culture (Livingstone & Russo, 2018). 3 females and 3 males were chosen at random of no particular age. The audio clip consists of 2 spoken sentences and are of length around 3 seconds. Since the 3 cultures were procured from various sources, there was a need for having common sampling and bit rate. Each dataset is easily available on the Internet and is free to use. A more detailed description of the datasets is given in Table 2 below:

Table 2. Database Description

Dataset	**Male**	**Female**	**Emotion**	**Audio quality**	**Total number of sentences**
Canadian French	6	6	Anger, Disgust, Fear, Happiness, Sad, Neutral	48KHz, Mono	443
EMOVO Italian	3	3	Anger, Disgust, Fear, Happiness, Sad, Neutral	48KHz, Stereo	510
North American English	12	12	Anger, Disgust, Fear, Happiness, Sad, Neutral, Calm	48Khz, Mono	1440

For our project, the 4 emotions that are being considered are: *happy*, *angry sad*, and *neutral*. This was done to reduce the complexity of the model and to also bring the three languages to the same terms. Since the datasets were composed from different sources, they naturally had different number of samples. Therefore, in order to make them balanced, while using 4 emotions of the same kind, we had to use similar number of samples, hence we reduced the larger datasets as follows: 299 for Canadian French, 286 for Italian, 299 for North American English. But more training samples were needed for better performance of our network, therefore we used data augmentation techniques to achieve the same. Data augmentation techniques introduce subtle differences in the samples to make them look different to the system without compromising on the information we wish to extract from them (Dyk & Meng, 2001). In this work, we have used two augmentation techniques in order to triple our datasets. They were:

- *Pitch Tuning*: with this technique, we shifted the pitch of the audio file and therefore created an augmented datafile with different pitch characteristics, but with the expression of the same emotion as the original. For a computer, this is equivalent to a different actor expressing the same emotion. We kept both.
- *Addition of white noise*: this is another popular data augmentation technique where random noise is introduced to audio file. To the human ear the white noise is a constant background sound and is usually present in poorly recorded sounds. In its core, any white noise has a flat frequency spectrum over the

range of all present frequencies in the signal. After adding the noise we kept both samples again, just as we did before.

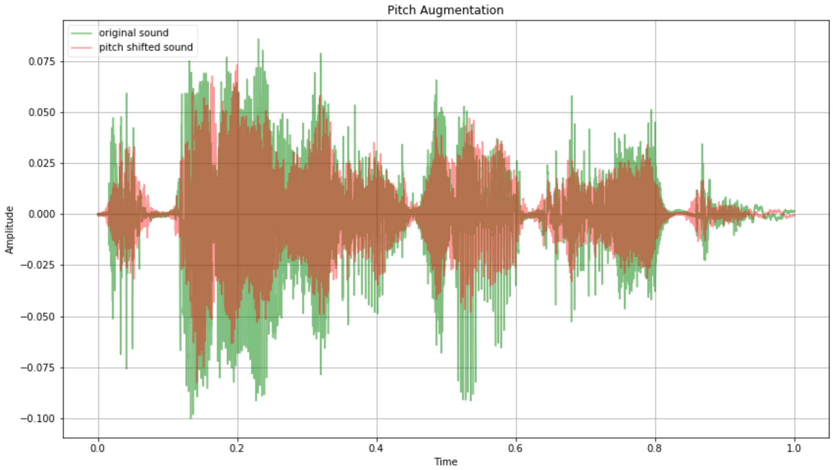

Fig. 2. Before and after pitch augmentation

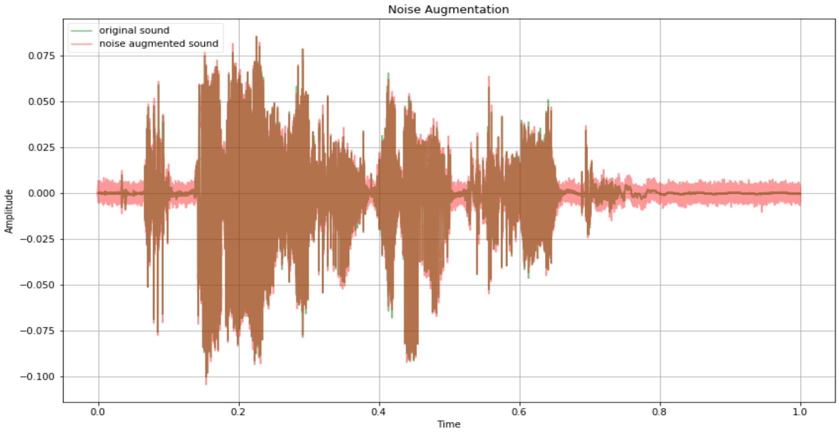

Fig. 3. Before and after noise augmentation

As can be observed from Figures 2 and 3, the changes we introduced to the original signal were significant and are visible in the plots. Tables 2 and 3 on the other hand display the number of speech samples before and after augmentation was performed.

Table 3. The number of samples before and after augmentation

Dataset	Number of samples before augmentation	Number of samples after augmentation
Canadian French	299	897
EMOVO Italian	286	858
North American English	299	897

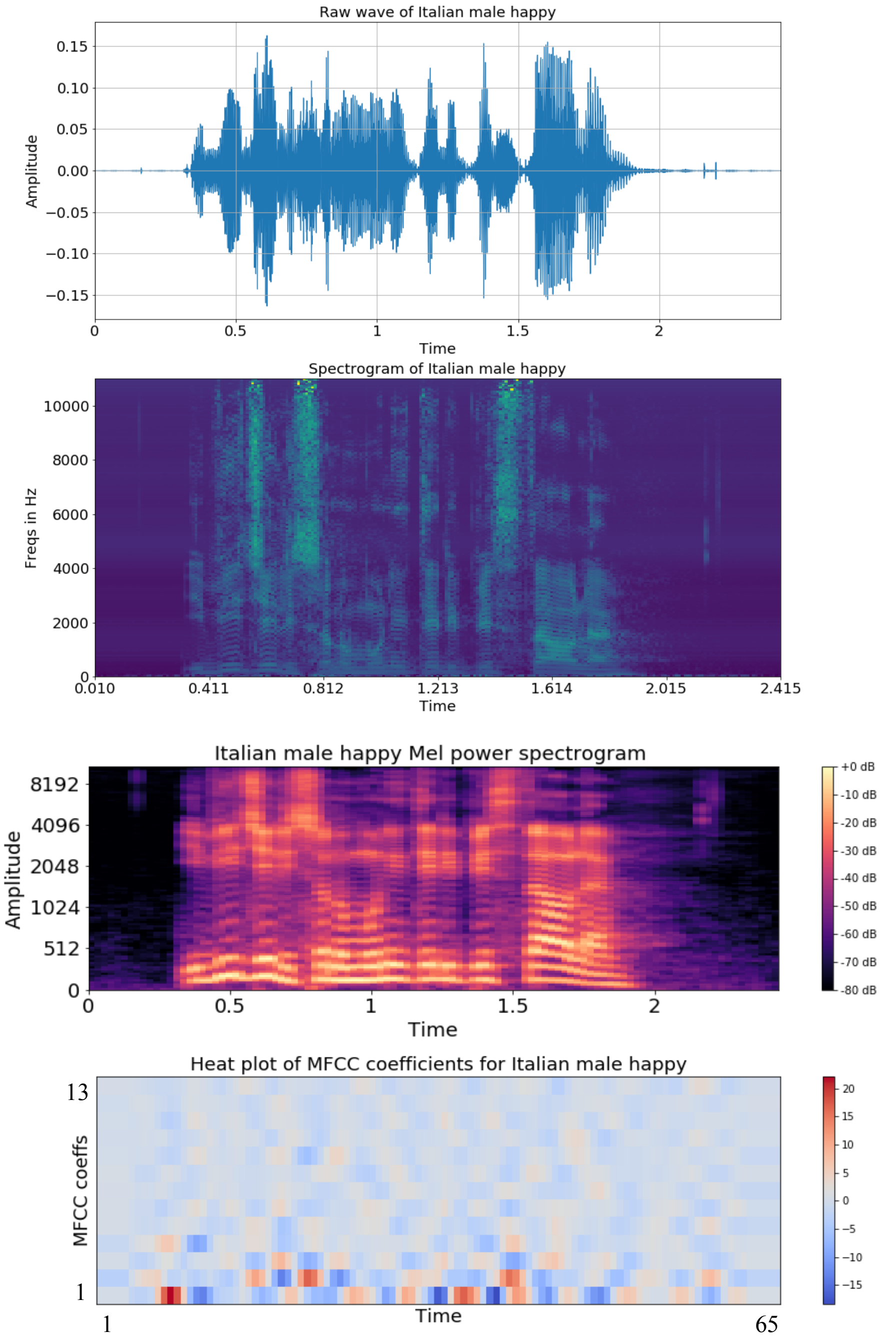

Fig. 4. Fig. 4a. Raw wave of Italian male happy;
Fig. 4b. Spectogram of Italian male happy; Fig. 4c..Italian male happy Mel power spectrogram;
Fig. 4c. Heat plot for MFCC coefficients for Italian male happy

After augmenting the data, we transformed the audio files into a form, from which we can extract the important information pertaining to the chosen set of emotions for this work. Audio files are often converted into the Mel frequency scale for different analysis. In more details, the raw audio file (represented in figure 4a) was first converted into the frequency spectrum using the Fast Fourier Transform. This transformed data is represented in figure 4b, where a spectrogram of the particular Italian sentence is given. We further transformed the file to the Mel scale and the Mel power spectrogram, and the result is depicted below in figure 4c. After transformation, we obtained the Mel frequency cepstral coefficient (MFCC), which we used for training out convolutional neural network. Figure 4d visualizes the variation of the coefficients through time for a single speech utterance, where the time of roughly 2.5 seconds is sectioned into 65 blocks, all of them depicted in a heat plot. We calculated 13 MFCC's for our purposes as shown on the *y-axis* and time was split into 65 slots as shown on the *x-axis*.

4.2 Neural Network Description

Training and Testing our model with the MFCCs we chose for each timeframe, was needed in order to verify that these features can be used to compare the different emotions for each of the three cultures for both genders. We trained a Neural Network, which provided an elegant solution to the complex non-linear problem we faced. In particular, a Convolutional Neural Network was found to be relevant to use for our system. We trained the system to give us an informed supervised decision using the combined and augmented training set we prepared in the previous step. We used several convolutional layers in order to achieve better accuracy. Figure 5 describes the Convolutional Neural Network that was used in this project. Using the MFCCs, we were able to extract 65 features from each utterance (speech sample) (Chu, 2019) and this feature vector was sent to the input of our CNN:

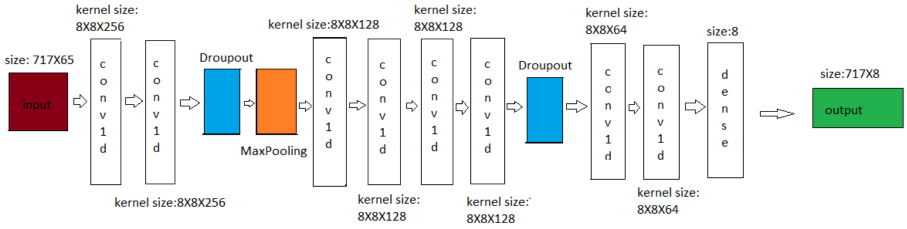

Fig. 5. Convolutional Neural Network Description

The following describes the model in much more detail:

1 Conv1d – the total number of one-dimensional convolution layers was experimentally chosen to be 8. In more complex dataset it is well known that convolution is usually improving the solution to the heavily non-linear problem.

Therefore, it was found that these numbers of layers was sufficient for the task at hand.

2 ***Activation*** – There were a total of 8 activation functions, one for each *conv1d* layer, out of which 7 were *ReLU* and 1 for the output was *Softmax*. They are used for essentially switching on or off the neural network layer as per demand.

3 ***Batch_normalization*** – it helped in accelerating the learning of our model by allowing higher learning rates. This is important, if the samples have been obtained from different sources that have various distributions. *Batch_normalization* helps by making the inputs stable and thus, making the neural network learn faster. In our case we had three different datasets, which we combined into a single dataset, hence we added *batch_normalization* after the second layer.

4 ***Dropout*** – overfitting is a huge problem that is most definitely made when having poorly designed neural networks. *Dropout* is a kind of a regularization method that prevents exactly this by randomly switching off some of the neurons or by 'dropping them out' so as to mimic real life in which we forget and force to learn (adjust the weights) again. We added 2 *dropout* layers, one after the 2nd *conv1d* and one before the 7th *conv1d* layer.

5 ***MaxPooling*** – is a dimensionality reduction technique used primarily for reducing overfitting. *MaxPooling* reduces the number of samples and gives a shortened representation of the feature sets in each step. This helps in reducing the computational cost by reducing the number of features to learn. We used *MaxPool* twice, once immediately after the first *dropout* before layer 3 and once after *activation_6*, where we set the *dropout* to be on ¼ (or 25%) of the feature set.

6 ***Flatten*** – it pools the whole matrix into a single column vector and then feeds it to the network for further processing. So in our case we arrived at a vector of 65 features.

7 ***Dense*** – the fully connected neural network layer was connected to the next layer with the help of a dense function.

The 2 activation functions used for this Model were:

1 ***Relu*** – when a model uses Stochastic Gradient Descent with backpropagation it requires a nonlinear activation function. Rectifier linear unit (*ReLU*) does exactly that by obtaining the maximum unit from a given range, or:

$$Y = Max(0, x) \tag{3}$$

2 ***Softmax*** – it is mostly used at the output layer as it sets out to give probabilities of the output. It is heavily used in multi-class classification. In this model it helps in recognizing the different emotions across cultures and genders. In our case we had 8 probable outputs one for each emotion (4) and gender (2).

As for the hyperparameters, the number of epochs chosen was 700. The learning rate was 0.0001. Stride was set to equal to 1 by default. Kernel's row and column size was kept at [8x8] respectively throughout the network, but the depth was reduced from 256 to 64 as the network progressed.

An important step while training a neural network is splitting the available dataset into training and testing sets. We used stratification shuffle technique for this purpose as it preserves the proportion of emotions in the parent set and maintains it across the training and testing set (Sechidis, Tsoumakas, & Vlahavas, 2011). Tables 4, 5 and 6 provide more details:

4.3 Testing and Training Samples

Table 4. French Train and Test Samples

Emotion	Male	Female	Train	Test
Happy	216	216	173(Male), 173(Female)	43(Male), 43(Female)
Anger	108	108	86(Male), 86(Female)	22(Male), 22(Female)
Sad	21	12	17(Male), 10(Female)	4(Male), 2(Female)
Neutral	108	108	86(Male), 86(Female)	22(Male), 22(Female)

Table 5. Italian Train and Test Samples

Emotion	Male	Female	Train	Test
Happy	84	126	67(Male), 101(Female)	17(Male), 25(Female)
Anger	84	126	67(Male), 101(Female)	17(Male), 25(Female)
Sad	84	84	67(Male), 101(Female)	17(Male), 25(Female)
Neutral	102	126	81(Male), 101(Female)	21(Male), 25(Female)

Table 6. North American Train and Test Samples

Emotion	Male	Female	Train	Test
Happy	123	120	98(Male), 96(Female)	25(Male), 24(Female)
Anger	90	102	72(Male), 82(Female)	18(Male), 20(Female)
Sad	93	105	84(Male), 84(Female)	19(Male), 21(Female)
Neutral	132	132	105(Male), 106(Female)	27(Male), 26(Female)

5 Results and Discussion

The results were plotted in Jupyter notebook using Pandas Library and are shown in Figure 6. There was a difficulty in getting prior results where we did not use data augmentation. The original accuracy before data augmentation was 53%, which was

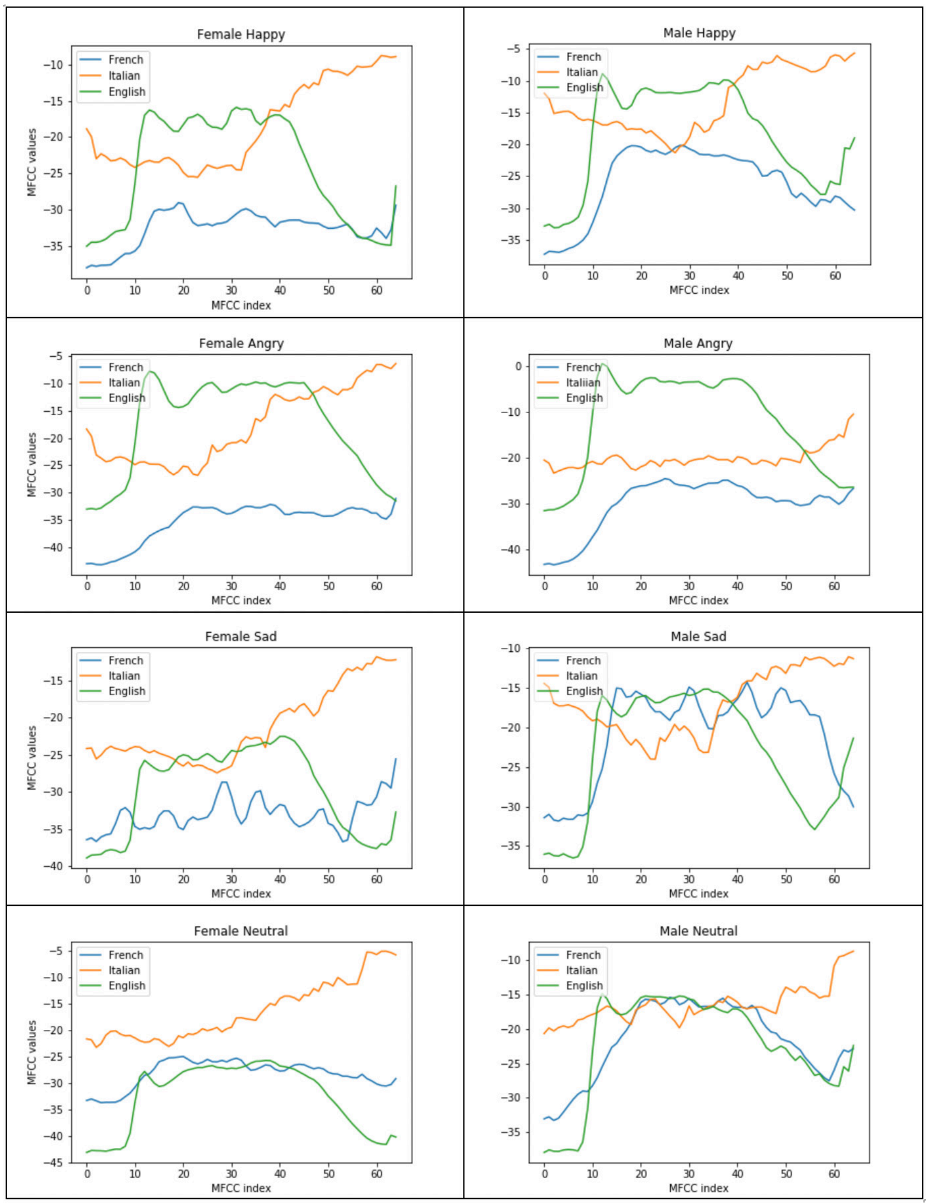

Fig. 6. Results

not good enough. This showed that the model was ineffective in deducting any kind of results since the number of samples was less than what we needed. We therefore decided to use data augmentation to enlarge dataset and improved our accuracy greatly. As mentioned already, we tripled our dataset using this technique and found that the network trained on this dataset performed significantly better for all three languages. We achieved an accuracy of 71.10 percent for Canadian French, 79.07 percent for Italian dataset and 73.89 for the North American dataset. This suggested that: 1) the initial low accuracy was due to insufficient size of our original datasets, and 2) that the MFCC parameters were effective features to capture different emotions in different cultures and genders.

Focusing on the result depicted in figure 6, we need to mention that each point on the plot represents the mean of the individual MFCC coefficients taken across samples for individual cultures. In the first section, as can be seen from the graph, for the 2 genders, emotion *happy* is expressed differently by the three cultures, the French, the Italian and the North American. Similarly, it is the case for the emotion *angry*, and it is clear that the three cultures have different ways of expressing these emotions between the two genders. But a break is seen in this trend as we come to the emotion *sad.* The *female* category has similar traits for the Italian and English culture as they overlap at some MFCC values. Similar observation can be drawn for the *male* section for the *sad* emotion where we can see some similarity among the three cultures. At the final set of plots, that is the *neutral* emotion, for the *male* category in all corresponding cultures an overlapping can be observed stating that there is a high level of similarity of expressing this emotion across the three cultures. In the *female* category, North American and the French culture can be seen as overlapping suggesting that they are much more like each other than the Italian culture. Overall, it is observed that the *happy* and *angry* emotions were expressed with more energy by every gender across all cultures.

6 Conclusions and Future Work

Expressing and understand emotions across different cultures have long been a problem for many people and have made them face difficult challenges. This paper tries to make that challenge somewhat easier by comparing the emotions across three different cultures, while validating the parameters through the use of Neural Networks. It can easily be seen how these cultures are different and sometimes similar with respect to one another. In the future, a much more detailed work can be expected using more cultures in order to gain a basic understanding of how each culture operates so that we can make more informed decisions. Further investigation is needed as to determine how the pitch shift and white noise played a role in extending the datasets and how they really diversified the number of new samples in order to increase the accuracy in each of the three models.

Acknowledgments

This work was partially supported by the Bulgarian Ministry of Education and Science under National Scientific Program “Information and Communication Technologies for a Single Digital Market in Science, Education and Security”, approved by DCM No 577, 17 August 2018.

References

Kwonl, O.-W., Chan, K., Hao, J., & Lee, T.-W. (2003). Emotion Recognition by Speech Signals. *8th European Conference on Speech Communication and Technology* (pp. 125-128). Geneva: EUROSPEECH 2003 - INTERSPEECH 2003.

Owens, A. (2008). A Case study of cross cultural communication issues for Filipino call centre staff and their Australian customers. *2008 IEEE International Professional Communication Conference* (pp. 1-10). Montreal: IEEE.

Iliev, A. I., & Stanchev, P. (2017). Smart Function Digital Content Ecosystem using Emotion Analysis of Voice. *International Conference on Computer Systems and Technologies.*

Livingstone, S. R., & Russo, F. A. (2018, May 16). *Plos One.* Retrieved from NCBI: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5955500/

Costantini, G., Iaderola, I., Paoloni, A., & Todisco, M. (2014). Emovo Corpus: an Italian Emotional Speech Database. *International Conference on Language Resources and Evaluation*, (pp. 3501-3504).

Cowen, A. S., & Keltner, D. (Sept, 05, 2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proceedings for the National Academy of Sciences of the United States of America.

Iliev, A. I., & Scordilis, M. S. (2008). Emotion Recognition in Speech using Inter-Sentence Glottal Statistics. *Proceedings of the 15th International Conference on systems, Signals and Image Processing* (pp. 465-468). Bratislava, Slovakia: IEEE-IWSSIP.

Iliev, A. I., & Scordilis, M. S. (2011). Spoken Emotion Recognition Using Glottal Symmetry. *EURASIP Journal on Advances in Signal Processing. 2011*, p. 11. Hindawi Publishing Corporation.

Iliev, A. I., & Stanchev, P. L. (2018). Glottal Attributes Extracted from Speech with Application to Emotion Driven Smart Systems. *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management* (pp. 297-302). KDIR.

Iliev, A. I. (2012). *Emotion Recognition From Speech.* Lambert Academic Publishing.

Gournay, P., Lahaie, O., & Lefebvre, R. (2018). A canadian french emotional speech dataset. *9th ACM Multimedia Systems Conference* (pp. 399-402). Amsterdam: Association of Computing Machinery.

Dyk, D. A., & Meng, X.-L. (2001). The Art of Data Augmentation. *Journal of Computational and Graphical Statistics* (pp. 1-50). Journal of Computational and Graphical Statistics.

Chu, R. (2019, June 1). *Speech Emotion Recognition with Convolutional Neural Network*. Retrieved from towardsdatascience: https://towardsdatascience.com/speech-emotion-recognition-with-convolution-neural-network-1e6bb7130ce3.

Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the Stratification of Multi-label Data. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 145-158). Lecture Notes in Computer Science, vol 6913.

Received: June 16, 2020
Reviewed: June 27, 2020
Finally Accepted: July 05, 2020