CSE527 Course Project Final Report
— Bimodal Regression Model for Personality Inference in Making First Impression

Zhenxun Zhuang, 111131765
Stony Brook University
Stony Brook, NY, US
zzhuang@cs.stonybrook.edu

Abstract

With every new encounter, one is assessed and another person's impression of him is formed. The first impression could greatly affect or even decide how one will be treated afterwards; thus, it is important to figure out how to make a good first impression. Among all factors that can influence the impression one makes, personality plays the most important role. Although not directly observable, personality can be inferred from an individual's expressions, gestures, and other external cues. Humans are very skilled and efficient in this that a quick glance is usually enough. Inspired by that, we strive to enable machines to automatically generate personality inferences. By providing instant feedback of how one will be viewed by others, this work can help people present themselves better by changing their behavior in simple ways. In order to exploit information from both visual and audio channels just like humans do, studies are carried out along two lines. For the visual modality, a modified CNN is employed, while a logistic regressor is used in the audio part. These two parts are then ensembled to obtain final predictions. Extensive research has been done to further boost performance, which leads to a 88.3% accuracy in the sense of Mean Absolute Accuracy.

1. Introduction

"You never get a second chance to make a first impression," a saying famously goes. The first impression is highly important in many contexts, such as dating, human resourcing, or job interviews. Its importance can never be too emphasized in that the first impression sets the tone for all the relationships that follow. Moreover, once a negative first impression is formed, it will take a lot of time and efforts to reverse it. The question is, how can we make a good first impression?

To answer that, we must find out what affect the impression we left on others. Possible influential factors include expressions, gestures, clothing, and so on. In fact, they are all external manifestations of one core property – personality. Personality refers to stable individual differences in characteristic patterns of thinking, feeling and behaving. It affects every aspect of one's daily life, from his clothing preferences and eating habits, to the composition of his social circle. Therefore, when we are gaining the impression of someone, we are actually trying to infer his personalities. In light of this, personality analysis was introduced to reveal the relation between people's behavior and the impression they give to others, thus help them present themselves better by changing their behavior in simple ways.

Personality analysis has become the subject of scientific inquiry ever since ancient Greek, and it remains to be a hot topic in modern psychology. It is of great interest due to its effectiveness in explaining and predicting observable behavioral differences, such as physical and psychological health, community involvement, and criminal activity [1].

Since personality is an abstract psychological construct which cannot be directly observed, how can we measure it? The dominant model used in personality analysis is trait model, which builds upon human judgments about semantic similarity and relationships between adjectives that people use to describe themselves [2]. It represents personality in terms of numerical values, a form particularly suitable for computer processing. After several decades of trial and error, the Big Five (BF) traits or Five-Factor Model (FFM) has been shown to *"provide a set of highly replicable dimensions that parsimoniously and comprehensively describe most phenotypic individual differences"* [3]. It models the human personality along five dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness.

Humans are very skilled at inferring personalities. Usually, we can get a rough impression even from just a glimpse. Inspired by that, the computing society strives to study how to enable machines to automatically recognize or synthesize personality traits. Numerous methods have been proposed, but the problem remains unsolved. Challenges come from many aspects, for instance, cultural and individ-

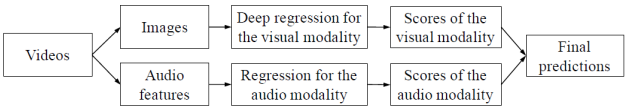


Figure 1: Framework of the Deep Bimodal Regression method.

ual differences in tempos and styles of articulation, variable observation conditions, and infinitely many kinds of out-of-vocabulary motion.

In most cases, one infers others' personalities and forms the first impression by both observing their behavior, and listening to their utterances. Therefore, it is natural to treat the problem along two lines, i.e., the visual and the audio modalities. In these two modalities, neural networks can be employed for capturing both visual and audio information which are then ensembled to obtain the final predicted personalities. This framework is called the Deep Bimodal Regression (DBR) framework, and was proposed in [4]. It is shown in Figure 1.

During the project, the major problem I met is the lack of enough computing resources, thus I paid much attention to how to reduce the complexity. The most significant breakthrough was achieved in studying to what extent can we decrease the number of frames extracted from each video while maintaining the performance. The result is rather surprising that even if only one frame is extracted per video, our model is still able to converge to the same performance. This suggests that, just like humans, personality traits inference can be obtained in very short exposure time. This finding is the cornerstone of the following research.

Observing that one's expressions, gestures, and behavior are in different scales, from involuntarily biting his bottom lip suggesting lying, to a twist of the whole body showing unease, I studied the relation between the performance and the downsampling factor of input images. The result shows that the best performance is achieved when input images are downsampled to a medium size. This means that both details and global information contribute to the accurate inference of personalities, which is in accord with our daily experience.

For the audio modality, I did extensive research on finding the audio representation that best fits this task. Not surprisingly, Mel Frequency Cepstral Coefficients (MFCC) which represents more stable characteristics of audios outperforms others since we pay less attention to transitory features when inferring personalities.

Combining all these techniques, our model achieves an accuracy of 88.3% in the sense of Mean Absolute Error (MAE).

2. Related work

2.1. Personality Computing

Although personality is an abstract psychological construct not directly observable, it can be perceived by externalizing through *distal cues*, i.e., traces of personality appearing in virtually everything observable an individual does. Distal cues reaching an observer go through a perception process which results into mental representations of things being perceived. Only certain cues, called *proximal cues*, are actually perceived by the observer, and they activate the attribution process which generates perceptual judgments of the personality of the individual being observed.

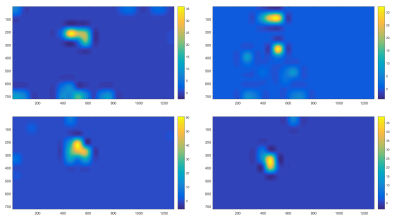
Three main problems are addressed in personality computing, namely Automatic Personality Recognition (APR), Automatic Personality Perception (APP), and Automatic Personality Synthesis (APS) [5]. APR strives to infer self-assessed personality from machine detectable distal cues, whereas APP aims to infer the personality observers attribute to an individual from proximal cues. However, methodologies adopted in APR are also effective in APP since they both try to find the relationship between cues and personality traits. On the other hand, APS is the task of automatically generating distal cues aimed at eliciting the attribution of desired personality traits by human observers.

2.2. Convolutional Neural Networks

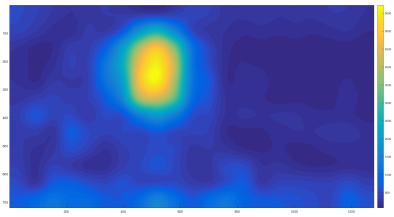
Though originally inspired by the goal of modeling biological neural systems, the area of artificial neural networks has since diverged and evolved into an independent subject. As suggested by its name, a neural network is a collection of neurons. What makes it unique is that neurons are organized into several distinct ordered layers instead of an amorphous blob. Cycles are not allowed, and usually only neurons in adjacent layers can be connected. The most common layer type is the fully-connected layer in which all neurons in one layer are pairwisely connected to every neuron in its preceding layer.

When the input of neural networks are images, several explicit assumptions can be made based on properties of images. Since in most cases features in images only occupy a small region, it is reasonable to confine the connectivity range, aka the receptive field, instead of using a fully connected layer. Moreover, if one feature, such as edges and corners, is useful to compute at some spatial position, then it should also be useful to compute at a different position. Therefore, a same set of parameters can be shared throughout a layer. By introducing these two techniques, the relationship between two adjacent layers actually becomes a convolution, hence the name CNN. The major advantage of CNN is the substantially reduced number of parameters needed to be learned, which makes it possible to train com-

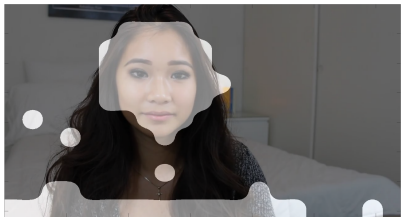
(a) Original image



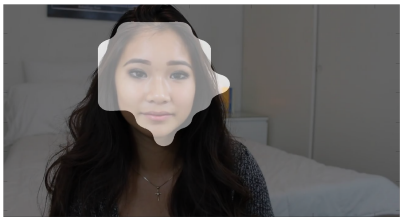
(b) Sample feature maps



(c) Activation map



(d) Significant set



(e) Mask

Figure 2: Pipeline of SCDA. ((b)-(e) are all upsamplesd to the same size as (a) for clarity.)

plex networks consisting of multiple layers to achieve good results in image-related tasks.

2.3. Audio Representations

Many representations have been proposed to identify the components of the audio signal, for example, Octave Band Signal Intensity (OBSI) [6], Line Spectral Frequency (LSF) [7], Linear Predictor Coefficients (LPC) [8] and Mel Frequency Cepstral Coefficients (MFCCs) [9].

Among them, MFCC are most widely used. It is a kind of short-term spectral-based features of an audio signal. A cepstrum is the Inverse Fourier Transform (IFT) of the logarithm of the estimated spectrum of a signal, thus it is also known as the spectrum of a spectrum. MFCC differs from the common cepstrum by incorporating the Mel scale that relates human perceived frequency, or pitch, of a pure tone to its actual measured frequency. Hence, MFCC matches more closely what humans actually hear.

3. The Bimodal Regression Model

Videos are comprised of images and accompanying sounds. In order to fully exploit the information carried by a video, both modalities should be used. The visual modality is addressed in Section 3.1, and the audio modality in Section 3.2. I will then explain how to ensemble these two parts in Section 3.3.

3.1. Visual Modality

Frame Extraction In order to utilize convolutional neural networks whose input is images, it is necessary to extract frames from provided videos. Since neighboring frames are usually similar, extracting all frames will be overkill. Moreover, existing computing resources cannot afford that. We will discuss to what extent can we decrease the sampling rate without deteriorating the performance in Section 4.1.

Feature Extraction Utilizing pre-trained CNNs to extract features from images has become a popular method. However, as shown in Figure 2(b), apart from the object that we are interested in, the output activation of convolutional layers also contains a lot of noise and background regions. Therefore, selecting and using only meaningful descriptors is necessary. To this end, the Selective Convolutional Descriptor Aggregation (SCDA) technique [10] was introduced.

SCDA modifies a pre-trained CNN (in this project, I use VGG-Face [11]) by discarding its fully connected layers. The activation of the last layer Pool_5 is instead processed as follows:

1. Sum the activation along the 3rd dimension to obtain the "activation map" $\mathbf{A}$ (as in Figure 2(c)) in which the higher activation response a particular position $(i; j)$ is, the more possibility of its corresponding region being part of the object.
2. Set the threshold as the mean value of $\mathbf{A}$, and choose those positions whose response is higher than the threshold to form a significant set $\mathbf{M}$ (as in Figure 2(d)).
3. Find the largest connected component in $\mathbf{M}$, which is denoted as $\widetilde{\mathbf{M}}$ (as in Figure 2(e)), to get rid of the interference of small noisy parts.

4. Use $\widetilde{\mathbf{M}}$ as the mask to select useful and meaningful descriptors from the original activation.
5. Max pool the selected descriptors, and follow by the standard $\ell 2$ normalization to get a 512-d representation.
6. Average pool the selected descriptors, and follow by the standard $\ell 2$ normalization to get a 512-d representation.
7. Concatenate these two 512-d representations to get a 1024-d representation of the input image.

Regressor For each trait of the Big Five model, two fully connected layers are built to exploit information in features. Since this is a regression task, the sigmoid function is employed to obtain prediction values. The $\ell 2$ distance is used as the loss function to calculate the regression loss.

3.2. Audio modality

Feature Extraction Since MFCC is the most widely used audio feature, I choose it as the audio representation. It can be computed as follows:

1. Segment the audio signal into short frames.
2. Calculate the power spectrum of each frame.
3. Apply the Mel filter banks to the power spectrum and sum the energy in each filter.
4. Take the logarithm of all filterbank energies.
5. Take the DCT of the log filterbank energies.
6. The MFCCs are the amplitudes of the first several (usually 13) DCT coefficients.

The formula converting from frequency to Mel scale is:

$$M(f) = 1125 \ln(1 + f/700) \tag{1}$$

To go back:

$$M^{-1}(m) = 700(\exp(m/1125) - 1) \tag{2}$$

Apart from the power spectral envelope of each frame described by MFCC, information is also included in the dynamics, i.e. the trajectories of the MFCC coefficients over time. To calculate the delta coefficients, the following formula is used:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{3}$$

where d_t is the delta coefficient in frame t computed from the static MFCC coefficients c_{t-N} to c_{t+N}.

Logistic Regression Similar to the visual modality, a logistic regressor is trained using the extracted features, and the $\ell 2$ distance is used as the loss function to calculate the regression loss.

3.3. Ensemble

After obtaining prediction values from both visual and audio modalities, they are averaged to get final personality inferences.

4. Experiments

Dataset The ChaLearn Looking at People First Impressions Challenge dataset consists of 10,000 15-second clips extracted from more than 3,000 different YouTube high-definition (HD) videos of people facing and speaking in English to a camera [12]. The people appearing are of different gender, age, nationality, and ethnicity, which makes the task of inferring apparent personality traits more challenging. The dataset is divided into three subset: training (6000 clips), validation (2000 clips), and test (2000 clips). The ground truth is obtained using Amazon Mechanical Turk. Since only the training set ground truth has been released, I further divided the training set into three subsets: training (3600 clips), validation (1200 clips), and test (1200 clips). Experiments in Section 4.1 and 4.2 are carried out on the training set, and evaluated on the validation set. In Section 4.3 the final performance of our model obtained from the test set will be reported.

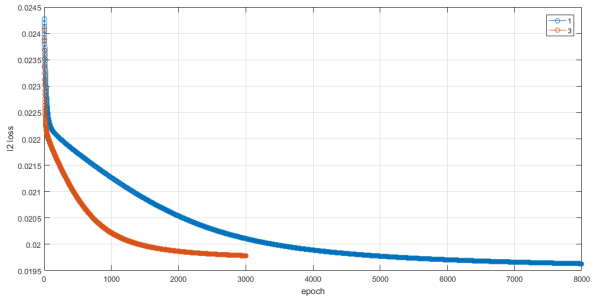
Evaluation The evaluation consists in computing the Mean Absolute Accuracy over all traits and videos:

$$\text{Mean Accuracy} = \frac{1}{5N}\sum_{j=1}^{5}\sum_{i=1}^{N} 1 - |gt_{i,j} - pred_{i,j}| \tag{4}$$

where N is the number of videos, $gt_{i,j}$ stands for the jth personality trait of the ith $ground\ truth$, and $pred_{i,j}$ stands for the jth personality trait of the ith $predicted\ value$.

4.1. Visual Modality

Effects of Adding More Layers Considering that one fully connected layer may not be enough to exploit all information in the extracted frames, I studied the effect of adding more layers. In this experiment, 10 frames are extracted from each video. Each added layer is a fully connected layer consisting of 256 nodes. From figure 4, we can find that more layers do improve performance, but the learning curve will start to distort. Since parameter number and training time both increase rapidly when increasing layers, and more layers will lead to overfitting, I adopt the two-layer model.



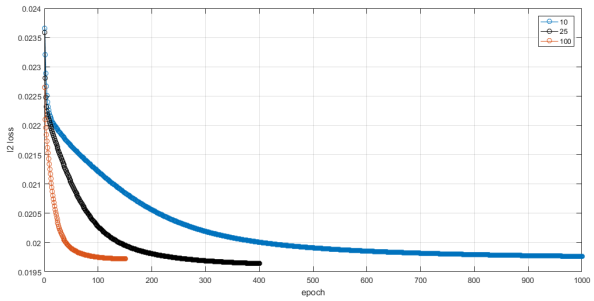


Figure 3: Comparison of learning curves (validation) when different number of frames are extracted per video.

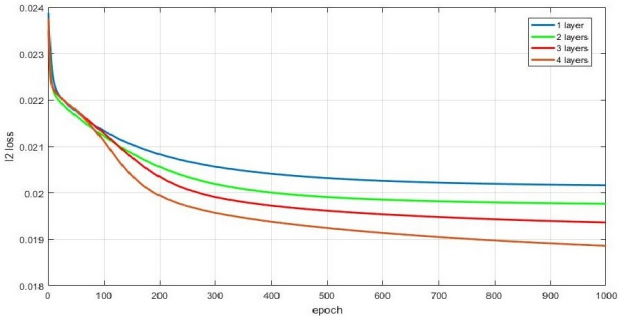


Figure 4: Comparison of learning curves (validation) when different number of fully connected layers are added after the SCDA model.

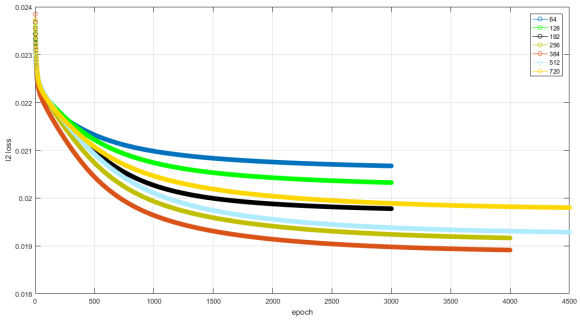


Figure 5: Comparison of learning curves (validation) when input images are downsampled to different sizes (numbers are heights and the ratio of height to width is preserved).

Number of Extracted Frames per Video Each video is approximately 15 seconds long with a frame rate of 30fps, thus there are about 450 frames in each video. Obviously, if we extract more frames from each video, we will get a larger training set. However, the time consumed in feature extraction will increase exponentially. Hence, it is of concern to find the smallest possible sampling rate above which more frames will not further boost the performance. I thus carried out the following experiment using the two fully connected layer SCDA model. Figure 3 shows that all networks achieve the same performance. Moreover, by multiplying the number of epochs and the number of extracted frames per video, we can find out that all networks actually converge at the same speed. Therefore, we can conclude that for this task, one frame per video is enough.

This result means personality traits can actually be inferred by only one static image, and dynamic videos do not carry much more useful information. Though quite surprising, it is indeed consistent with [13] which proves that even an exposure time as short as 100ms is sufficient for humans to make a personality inference from facial appearance, and additional exposure time will increase confidence in judgments but won't change the initial inference. Since human faces occupy the major area in all videos of this task, and the scenes do not change a lot, it is possible to infer personality traits from very few frames. Nevertheless, when the number of layers increases, or when the content of input videos changes relatively fast, more frames should be extracted from each video to provide enough training data.

Global Characteristics or Details In the VGG-Face model, frames are first downsampled before being fed into the network. A large downsampled image preserves many details, while a small one only contains global characteristics. In order to make out which scale is more important to personality inference, I compared 7 different downsampling factors, and respective heights of downsampled images are 64, 128, 192, 256, 384, 512, and 720 (the original ratio of width to height $1280 : 720$ is preserved). Since a large downsampled frame has many details which tend to change much faster than global features, I set the sampling rate to 3 frames per video. The result in Figure 5 indicates that when frames are downsampled to a medium size in which both details and global features are contained, our model works best. This means that both global and local characteristics are important for obtaining an accurate personality inference. This corresponds to our personal experience that when inferring personalities, we don't want to lose details, but we don't want to be overwhelmed by details, either.

Feature	Mean Accuracy	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Log_Filter_Bank	0.8806	0.8792	0.8934	0.8738	0.8770	0.8797
MFCC	0.8840	0.8835	0.8953	0.8767	0.8809	0.8835
Delta_MFCC	0.8807	0.8790	0.8935	0.8738	0.8771	0.8800
Comb_MFCC	0.8840	0.8835	0.8953	0.8767	0.8809	0.8835
Comb_MFCC_sn	0.8841	0.8836	0.8953	0.8768	0.8810	0.8836

Table 1: Comparison of different kinds of audio features. (Comb_MFCC means appending original delta_MFCC to the end of original MFCC and then $\ell 2$ normalizing the whole vector thus preserving the relative scale of these two features, whereas Comb_MFCC_sn means appending $\ell 2$ normalized delta_MFCC to the end of $\ell 2$ normalized MFCC.)

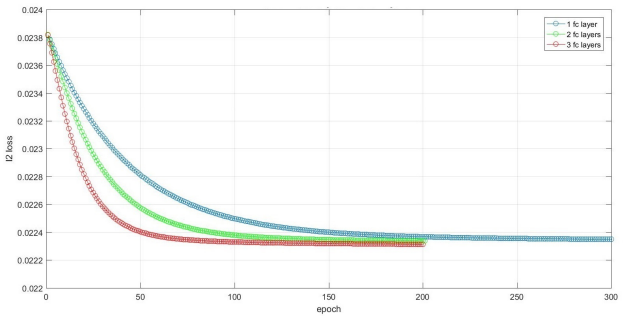


Figure 6: Comparison of learning curves (validation) when different number of fully connected layers are added in audio modality

4.2. Audio Modality

The FFmpeg[1] library is used to extract sound tracks from videos and transform them to the *wav* format. After that, Log Filter Bank Coefficients (LFBC) (26-d), MFCC (13-d), delta MFCC (13-d) are computed using a Python library *python_speech_features*[2]. In order to ensure features of all videos are of same size, periodic extension is employed. Specifically, since most audios have the same duration, 15 seconds, features extracted from those shorter audios will be padded in the end using elements from the beginning.

Comparison of Different Features In order to select the audio representation that best fits this task, I compared the performance of several features. Results in Table 1 shows that MFCC outperforms LFBC, and the combination of MFCC and delta MFCC slightly improves the performance. However, MFCC converges 10x slower than LFBC. This can be explained by their relationship: MFCC is the first a few DCT coefficients of LFBC, thus it only contains low frequency information of LFBC and discards high frequency one. Therefore, the diversity of MFCC coefficients is lower than that of LFBC which leads to slower convergence. However, since MFCC represents more stable features of signals, they can better model human voices whose characteristics change relatively slowly. Consequently, Comb_MFCC_sn is chosen to be the audio feature.

[1]http://ffmpeg.org/

[2]https://github.com/jameslyons/python_speech_features

Item	Mean	Extra.	Agree.	Consc.	Neuro.	Open.
MAA	0.8829	0.8827	0.8946	0.8757	0.8791	0.8822
RMSE	0.1447	0.1446	0.1312	0.1519	0.1489	0.1459

Table 2: The performance of our model evaluated on the test set (MAA = Mean Absolute Accuracy, RMSE = Root Mean Square Error).

Effects of Adding More Layers Due to the same concern as in visual modality that one fully connected layer may not be enough, I studied the effect of adding more layers. Each added layer is a fully connected layer consisting of 1000 nodes. Figure 6 shows that adding more layers can accelerate convergence, but it only provides marginal improvements. Meanwhile, when the network contains more and more layers, the number of parameters will increase dramatically and the learning process will be very slow. Therefore, only one fully connected layer is used.

4.3. Ensemble

The personality trait inferences obtained from visual and audio modality are averaged to be the final predictions. The final results evaluated on the test set are listed in Table 2. Since the results in the original paper are obtained on a different dataset, they are not listed here.

5. Conclusion

In this project, I try to learn a model that is able to automatically infer personality traits from human-centered videos. The combination of a modified CNN and SCDA followed by several fully connected layers is used to exploit visual information, while a logistic regressor is adopted in the audio modality. Beyond the DBR framework proposed in the original paper, I explored many possible directions to further improve the model, and achieved good results. However, due to the time limit, I am not able to study many other aspects. For example, the modified CNN used to extract features can be trained together with following fully connected layers.

References

[1] D. J. Ozer and V. Benet-Martinez, "Personality and the prediction of consequential outcomes," *Annu. Rev. Psychol.*, vol. 57, pp. 401–421, 2006.

[2] L. R. Goldberg, "The structure of phenotypic personality traits.," *American psychologist*, vol. 48, no. 1, p. 26, 1993.

[3] G. Saucier and L. Goldberg, "The language of personality: Lexical perspectives," *The five-factor model of personality: Theoretical perspectives*, p. 21, 1996.

[4] C.-L. Zhang, H. Zhang, X.-S. Wei, and J. Wu, "Deep bimodal regression for apparent personality analysis?," in *ChaLearn Looking at People Workshop on Apparent Personality Analysis, ECCV Workshop proceedings, p. in press. Springer Science+ Business Media*, 2016.

[5] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[6] S. Essid, *Classification automatique des signaux audiofréquences: reconnaissance des instruments de musique*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2005.

[7] T. Bäckström and C. Magi, "Properties of line spectrum pair polynomialsa review," *Signal processing*, vol. 86, no. 11, pp. 3286–3298, 2006.

[8] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[9] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.

[10] X.-S. Wei, J.-H. Luo, and J. Wu, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *arXiv preprint arXiv:1604.04994*, 2016.

[11] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *British Machine Vision Conference*, vol. 1, p. 6, 2015.

[12] V. Ponce-Lopez, B. Chen, M. Oliu, C. Cornearu, A. Clapes, I. Guyon, X. Baro, H. J. Escalante, and S. Escalera, "ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results," in *European Conference on Computer Vision (ECCV 2016) Workshops*, (Unknown, Unknown or Invalid Region), 2016.

[13] J. Willis and A. Todorov, "First impressions making up your mind after a 100-ms exposure to a face," *Psychological science*, vol. 17, no. 7, pp. 592–598, 2006.