Implementing Fusion Techniques for the Classification of Paralinguistic Information

Bogdan Vlasenko\(^1\), Jilt Sebastian\(^{1,2,3}\), D S Pavan Kumar\(^{1,3}\), Mathew Magimai-Doss\(^1\)

\(^1\)Idiap Research Institute, CH-1920 Martigny, Switzerland
\(^2\)Indian Institute of Technology Madras, India
\(^3\)École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

bo
dan.vlasenko@idiap.ch, jiltsebastian@gmail.com, dspavankumar@gmail.com, mathew.magimaidoss@idiap.ch

Abstract
This work tests several classification techniques and acoustic features and further combines them using late fusion to classify paralinguistic information for the ComParE 2018 challenge. We use Multiple Linear Regression (MLR) with Ordinary Least Squares (OLS) analysis to select the most informative features for Self-Assessed Affect (SSA) sub-Challenge. We also propose to use raw-waveform convolutional neural networks (CNN) in the context of three paralinguistic sub-challenges. By using combined evaluation split for estimating codebook, we obtain better representation for Bag-of-Audio-Words approach. We preprocess the speech to vocalized segments to improve classification performance. For fusion of our leading classification techniques, we use weighted late fusion approach applied for confidence scores. We use two mismatched evaluation phases by exchanging the training and development sets, and this estimates the optimal fusion weight. Weighted late fusion provides better performance on development sets in comparison with baseline.

Index Terms: fusion, feature selection, Multiple Linear Regression, raw-waveform CNN

1. Introduction

Computational paralinguistics covers various non-verbal information channels of speech and other types of vocalizations, and has developed rapidly over the last decade. The ComParE 2018 challenge comprises of four sub-challenges for predicting four basic emotions from the speech of handicapped subjects (Atypical effect), valence scores given by speakers themselves (Self-Assessed Affect (SSA)), three types of infant vocalizations (Crying) and three different types of heart beats (Heart Beats) [1]. The organizers of the challenge provide a baseline system composed of the state-of-the-art acoustic features and commonly used classification techniques. Feature extraction and machine learning techniques can be reproduced via freely available, open source tools [2,3]. It is difficult to outperform the baseline results owing to the complex fusion techniques used. A large variety of paralinguistic tasks are investigated in the recent years including but not limited to speaker traits [4], identifying child directed speech, cold and snoring identification [5], social signals, conflicts, emotions and autism [6].

As the paralinguistic applications tested varies from year to year, there has always been sufficient interest in feature selection and fusion techniques. Results presented in [7] show that combination of frame- and utterance level-analysis could significantly improve emotion recognition performance. Multi-modal decision level fusion is proposed in [8] for emotion recognition in the Wild [9]. Results presented in [10–12] show that the applied fusion methods improve the performance of the stand-alone detectors and provide systems capable of outperforming the baseline systems.

Feature selection for various models have been proposed in the past. This includes feature selection in Support Vector Machine (SVM) classifier [13] for emotion recognition, ensemble feature selection in [14] for model adaptation, and recent opensMI.LE [1]. Canonical correlation analysis (CCA) is employed in [15] for selecting apt features from a set of speech features for depression recognition. Acoustic, linguistic and psycholinguistic features are employed in [16] for learning the personality traits from the spoken data. Various feature selection methodologies in high-dimensional classification are presented in [17] where, speaker likability, intelligibility and personality traits are considered for classification. In [18], features extracted from a deep neural network are used for robust identification of child-directed speech, cold and snoring identification [9].

Year to year, the number of subjects/instances selected for the challenges has increased from few hundreds to several thousand. Increasing the number of acoustic features opens room for machine learning research, in optimizing feature set and improvement in robustness of classification techniques for high-dimensional fixed length turn-level features. Results presented during last ComParE challenges highlighted the importance of feature selection in handling high-dimensional paralinguistic datasets [19–21].

Recently, end-to-end systems are also used for classifying paralinguistic information. It avoids the use of hand-crafted features and allows the model itself to learn most suitable feature representation for the given task [5,22]. These applications are inspired by use of raw-waveform methods in ASR and speaker recognition tasks. We propose a raw-waveform CNN for three of the paralinguistic sub challenges. We also compare the proposed approach with the corresponding end-to-end baseline. As the confidence scores are not available for the baseline system [23], it is not possible to use it for fusion techniques.

The contribution of this paper is threefold. First, we propose to use Voice Activity Detector (VAD) for pruning irrelevant information located in silent segments of challenge datasets. Second, we propose to combine splits for estimation of codebooks for Bag-of-Audio Words approach introduced in [24]. Third, we propose a raw-waveform CNN which has a comparable performance with the end-to-end baseline and also provide reliable confidence scores at frame-level and turn-level for late fusion. We postulate that combined training, development and test splits for specification codebook representatives should improve classification...
2. General Framework

We present the general framework used for all the sub-challenges in this section. We remove the silence using a suitable VAD in the first step. Vocalized segments are used for acoustic feature selection using Low-Level-Descriptors (LLD) and functionals. We employed Multiple Linear regression with Ordinary Least Squares analysis, Bag-of-Audio-Words approach [24] with codebooks estimated on the combined splits, and CNN technique applied to raw waveforms for classification. Classification techniques applied varies for each sub-challenge. We finally fuse the scores from different best individual classifiers.

2.1. Speech Preprocessing

The preprocessing step consists of pruning irrelevant information using a suitable Voice Activity Detector (VAD). We used a Gaussian mixture model (GMM)-based VAD [25] which is more effective than simple energy-based counterparts when using with varying background noise levels. This converts the original signals to segments without silence. Finally, all non-silent segments were concatenated.

2.2. Acoustic Feature Extraction

For extraction of acoustic features, we used the openSMILE toolkit [26] with feature set configuration presented for Interspeech 2016 ComParE Challenge [27]. The same feature set has been used in the five previous editions of the Interspeech ComParE challenges as well. Original feature set contains 6373 static features resulting from the computation of various functionals over low-level descriptor (LLD) contours [28]. 65 LLDs and corresponding 65 delta coefficients are extracted for BoAW representation.

2.3. Modified Bag-of-Audio Words (BoAW) approach

In this work, we decided to used slightly modified BoAW approach for our experiments. Instead of using baseline configuration we decided to combine training, development and test split for better coverage of the codebook. Also, we skipped standardization step introduced in baseline system. We also used a concept of two codebooks: the first codebook is estimated for the 65 LLDs extracted from the ComParE feature set and the second codebook for the 65 deltas of these LLDs.

3. Methodology for SSA sub challenge

The dataset for Self-Assessed Affect sub-challenge comprises four times five-minute sessions from around 150 individuals. They had to speak spontaneously twice about a negative, and twice about positive experiences in their life. Before and after, the speaker reports a self-assessment of their own state of mind (Arousal and Valence on a ten-point Likert scale). In this challenge, the task is to determine the emotion of individuals as was assessed by themselves. We proposed to use the pipeline shown in Figure 1 for this challenge.

The speech signals are preprocessed using VAD to select vocalized segments. Feature extraction is performed by selecting LLDs and functionals as mentioned in Section 2. We use two classification techniques in this challenge; BoAW with SVM (Section 2) and multiple linear regression with ordinary least squares analysis.

Classification performance on the development set for BoAW approach with applied modifications is presented in Table 1. Confidence scores obtained with modified BoAW technique are marked as $P_{ij}(x_i)$ where $x_i$ is one of the possible classes.

3.1. Multiple Linear regression

For the selected acoustic features we implement least square analysis and estimate Multiple Linear Regression (MLR) coefficients. 28 linear separable features, presented in Listing 1 were selected from 6373 turn-level features for SSA sub-challenge task. Selected features contains various spectral features.
Classification performance on development set for MLR approach with selected features is presented in Table 1. Confidence scores for obtained with MLR technique are marked with $P_i(x_i)$, where $x_i$ is one of possible classes.

$$P_{\text{fusion}}(x_i) = \omega P_i(x_i) + (1 - \omega) P_I(x_i) \quad \forall x_i \in X$$ (1)

where $X$ is a set of all possible classes, $0 < \omega < 1$. Class $x_h$ with highest probability $P_{\text{fusion}}(x_h)$ is selected as recognized.

### 3.2. Late fusion

Weighted late fusion is applied for merging the confidence scores obtained with MLR and modified BoAW approaches. Equation (1) represents weighted fusion approach applied for fusion MLR (I) and BoAW (II) approaches.

**Table 1: UAR [%] rates for best baseline system and our approach. Results obtained on development set for Self Assessed Affect sub-challenge.**

<table>
<thead>
<tr>
<th>Case</th>
<th>System</th>
<th>UAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>Baseline</td>
<td>56.7</td>
</tr>
<tr>
<td>Direct</td>
<td>MLR</td>
<td>63.5</td>
</tr>
<tr>
<td>Direct</td>
<td>Reverse</td>
<td>60.1</td>
</tr>
<tr>
<td>Direct</td>
<td>BoAW(GLOB)+SVM</td>
<td>62.1</td>
</tr>
<tr>
<td>Reverse</td>
<td>BoAW(GLOB)+SVM</td>
<td>53.3</td>
</tr>
</tbody>
</table>

Listing 1: Selected features for Self-Assess Affect

<table>
<thead>
<tr>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>auSpec_Rfilt_sma [2]_percentile99.0</td>
</tr>
<tr>
<td>auSpec_Rfilt_sma [5]_lpc3</td>
</tr>
<tr>
<td>auSpec_Rfilt_sma [12]_lpc3</td>
</tr>
<tr>
<td>auSpec_Rfilt_sma [22]_upveltime90</td>
</tr>
<tr>
<td>pcm_fltMag_spectralEntropy_sma [2]_lpc4</td>
</tr>
<tr>
<td>pcm_fltMag_spectralVariance_sma,skewness</td>
</tr>
<tr>
<td>pcm_fltMag_spectralKurtosis_sma_quartilel</td>
</tr>
<tr>
<td>mfcc_sma [4]_kurtosis, mfcc_sma [8]_range</td>
</tr>
<tr>
<td>auSpec_Rfilt_sma [5]_lpc [3]_upveltime0-1</td>
</tr>
<tr>
<td>auSpec_Rfilt_sma [12]_kurtosis</td>
</tr>
<tr>
<td>auSpec_Rfilt_sma [20]_upveltime2</td>
</tr>
<tr>
<td>mfcc_sma [6]_lpc0, mfcc_sma [11]_lp gain</td>
</tr>
<tr>
<td>mfcc_sma [14]_maxPos</td>
</tr>
<tr>
<td>shimmerLocal_sma [pc4]_jitterDPP_sma,amean</td>
</tr>
<tr>
<td>pcm_RMSenergy_sma,siddevRisingSlope</td>
</tr>
<tr>
<td>auSpec_Rfilt_sma [6]_peakMeanMeanDist</td>
</tr>
<tr>
<td>mfcc_sma [5]_meanPeakDist</td>
</tr>
<tr>
<td>mfcc_sma [10]_peakMeanRel</td>
</tr>
<tr>
<td>mfcc_sma [12]_siddevRisingSlope</td>
</tr>
<tr>
<td>mfcc_sma [13]_qregc2, mfcc_sma [3]_flateness</td>
</tr>
<tr>
<td>auSpec_Rfilt_sma [7]_minRangeRel</td>
</tr>
<tr>
<td>pcm_fltMag_spectralFlux_sma,peakRangeAbs</td>
</tr>
</tbody>
</table>

Raw waveform methods have recently been exploited for speech processing applications with its inherent ability to extract features which are specific to the application. As a part of the challenge, we propose a raw-waveform CNN for Crying, Heartbeat and Self-assessed Assessment sub challenges and compare it with raw-waveform methods provided as the baselines. This is motivated by the use of raw-CNN for applications such as automatic speech recognition [29] and voice presentation attack detection [30].

The proposed network consists of two-to-three convolutional layers depending on the application and a hidden layer with relu activation function. This architecture of convolution layers followed by a multilayer perceptron is motivated by the success on various tasks, such as speech recognition [29, 31], presentation attack detection [30] and speaker recognition [32]. The output layer performs a softmax operation to obtain frame-based posteriors.

We provide the architecture of Baby Crying sub challenge for general understanding of the proposed approach. We use 400ms window length ($\text{W}_\text{len}$) with a 40 ms shift ($\text{W}_\text{shift}$) for making frame level decisions about the underlying classes. First convolutional layer has 50 filters ($\text{N}_\text{filter}$) with a filter width of 30 samples ($\text{N}_\text{shift}$) and is shifted by 5 samples ($\text{N}_\text{shift1}$) to have 75 such chunks over the window length. We use a max pooling size of 3 (3mp, $i = 1...N$) for all convolutional layers ($\text{N}$). Pooling is followed by non-linear activation and we choose relu based on its applications in ASR and ASV Spoof Detection tasks. Second convolutional layer has 60 filters ($\text{N}_\text{filter}2$), filter width of 7 ($\text{N}_\text{shift}$) and a filter shift of 1 ($\text{N}_\text{shift2}$). Hence, the total number of samples actually considered in filtering is $7 + 75 = 525$. Third convolutional layer (if exists) also has the same configurations as the second one. Parameters such as window length, number of filters, filter width and stride are selected based on heuristics and assumptions about the task. For example, a very small $\text{N}_\text{shift1}$ and ($\text{N}_\text{shift2}$) in Crying task is justified by the need of analysis at micro-level, so as to cover at least two pitch cycles of a baby, which is typically of 1 – 2 ms. The pitch plays a very important role in their voice as their vocal tract is not fully developed. A large $\text{W}_\text{len}$ captures the contextual information and it avoids the need of recurrent units.

In the challenge paper [1], development data is used for evaluating the top performing model and the testing is also performed with the same data. This could lead to a bias towards the development data and increase the performance gap with the testing data. As compared to the experiments reported in the challenge paper, we follow a different procedure to train the proposed and baseline networks. 90% of the training examples are used for training and rest is used for validation. Training instances are repeated in the training set so as to have an even distribution of instances among classes. The development data provided is used for testing the models. As the development data is unseen during the training, testing the model with this data is justified. The experiments are performed with 2 LSTM layers and 2 LSTMM layers as it has better unweighted average recall ($\text{UAR}$) in the development set [1]. The baseline model is trained for 10 epochs and stops training when the validation loss stagnates between successive epochs. Training is terminated when the LR drops below $10^{-5}$ and the final model is used for evaluation.
5. Results

During first experiment phase, we used development data for tuning classification parameters of MLR and modified BoAW techniques. During the second phase, we used development data for the estimation of late fusion weight. Afterwards, optimized architecture was used for training and fusing models on combined training and development sets. Finally, fused models were evaluated on test data.

5.1. Evaluation of development data

We used two different evaluation setups for SSA sub-challenge: we trained models on the training set and evaluated on the development set (direct case), and trained models on the development set and evaluated on the training set (reverse case).

One could see, that proposed MLR and BoAW with global codebooks outperformed results obtained with baseline system on development set. Afterwards, by using technique described in Section 3.2 we were estimating optimal $\omega$ value for late fusion of MLR and BoAW. Results obtained with direct and reverse evaluation cases were used. $UAR$ as a function of $\omega$ are presented in Figure 2.

![Figure 2: $UAR$ as a function of linear regression weight $\omega$. Direct case: training on training data, evaluation on development. Reverse order - training on development set and evaluation on training set.](image)

As one could see from Figure 2 curves for the direct and the reverse case have different maximum values. With $\omega = 0.426$ we can improve classification performance for the direct case from $UAR_{MLR} = 63.3\%$ and $UAR_{BoAW} = 62.1\%$ to $UAR_{fusion} = 67.0\%$. In the case of reverse case, on optimal fusion result was obtained with $\omega = 0.642$.

5.2. Evaluation of test data

By averaging $UAR$ curves for the direct and the reverse case we found that an optimal performance for fused technique can be obtained $\omega = 0.63$. Hence, for late fusion of MLR and BoAW techniques trained on combined training and development sets we used $\omega = 0.63$. Results obtained on test sets with proposed technique is presented in Table 3. The corresponding confusion matrix is given in Table 4, where we observe the highest confusion between the low (class $l$) and middle level (class $m$) of valence. As one could see recall rates for $m$ and $h$ valence classes significantly outperform $UAR$ reported for best baseline system. For final test evaluations we are planning to improve the results for valence class $l$.

5.3. Results on raw-waveform CNN

We compare the performance of proposed approach with raw-waveform-based END2YOU network \cite{23} for three sub-challenges. In terms of $UAR$, proposed raw-CNN approach has a comparable performance with baseline for Crying task, whereas it outperforms the baseline for other two tasks, see Table 3. This is achieved in spite of being a less complex network. For ex-

![Table 3: Confusion Matrix for evaluation on test set for Self Assessed Affect Sub Challenge.](image)

<table>
<thead>
<tr>
<th>System</th>
<th>Crying</th>
<th>Heart beat</th>
<th>Self-Assessed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>73.7</td>
<td>25.6</td>
<td>43.0</td>
</tr>
<tr>
<td>Proposed</td>
<td>71.1</td>
<td>44.8</td>
<td>49.2</td>
</tr>
</tbody>
</table>

ample, baseline system has 2,115,035 parameters in total as compared to 1,668,539 parameters of the proposed model for crying sub challenge. This is reflected in both training and testing times. The LSTM layers of the baseline causes delay in training the network and the models validation performance need not necessarily have a convergence with respect to the epochs.

6. Conclusions and Future work

In this work, we propose to use MLR for selecting acoustic features, and we offer to use advanced technology for BoAW codebook creation. We showed that implementing weighted late fusion for MLR and BoAW could significantly improve performance on the development set. We simulated direct and reverse evaluation cases for tuning the late fusion weight parameter for evaluating on the test set. The preliminary results on SSA development and test sets indicate that presented techniques are effective in the classification of Self-Assessed Affect. We also proposed a raw-waveform CNN for three sub-challenges. Raw-waveform techniques perform comparable to the baseline for Crying sub challenge and outperform the baseline results for SSA and Heart Beats sub challenge.

7. Acknowledgment

This work was partly supported by the IdeARK funded project COBALT, the Swiss Government Excellence Scholarship Project with ESKAS No: 2017.0575 and the HASLER Foundation project FLOSS.
8. References


