Speech Emotion Classification Using Tree-Structured Sparse Logistic Regression

Myung Jong Kim, Joohong Yoo, Younggwan Kim, Hoirin Kim

School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST)
myungjong@kaist.ac.kr, j9h2y1@kaist.ac.kr, cleanthink@kaist.ac.kr, hoirkim@kaist.ac.kr

Abstract

The extraction and selection of acoustic features are crucial steps in the development of a system for classifying emotions in speech. Most works in the field use some kind of prosodic features, often in combination with spectral and glottal features, and select appropriate features in classifying emotions. In the methods, feature choices are mostly made regardless of existing relationships and structures between features. However, considering them can be beneficial, potentially both for interpretability and to improve classification performance. To this end, a structured sparse logistic regression model incorporated with the hierarchical structure of features derived from prosody, spectral envelope, and glottal information is proposed in this paper. The proposed model simultaneously addresses tree-structured sparse feature selection and emotion classification. Evaluation of the proposed model on Berlin emotional database showed substantial improvement over the conventional sparse logistic regression model.

Index Terms: speech emotion classification, structured sparse model, logistic regression

1. Introduction

Automatic speech emotion classification, which aims to identify emotional states from speech signals, has been drawing increasing attention since it is able to make the human-computer interaction more natural and give further insights into emotional expression [1].

Each emotion induces physiological and psychological changes and these changes have certain influence on the speech characteristics [2], [3]. Therefore, research efforts in this field have mainly focused on feature extraction in the aspects of prosody, spectral envelope, and glottal waveform. Prosody is largely affected by the emotions and a fundamental cue for identifying emotions. Prosodic features including fundamental frequency, formants, and energy were widely exploited in emotion classification because they were related to the arousal level of emotions [4]-[6]. The vocal tract is also influenced by the emotional state, and therefore the spectral characteristics of the voice are changed depending on the emotions. Linear prediction cepstral coefficients (LPCCs) or Mel-frequency cepstral coefficients (MFCCs) were generally used as spectral features [7], [8]. Glottal flow is known to significantly vary with changes in phonation type, so it has an important contribution to the characteristics of speech under emotion [9]. As glottal features, normalized amplitude quotient [24], quasi-open quotient [25], and H1-H2 ratio [26] were widely employed.

Acoustic features of different nature are combined in order to increase the available information. For example, prosody is determined mainly by the vocal fold activity and the spectrum envelope is mostly influenced by the vocal tract, providing different information about the emotional state of the speaker. Therefore, a combination of different domain features may be important in obtaining better performances [10]-[12]. In recent years, research on feature selection has great attention to find the best feature set among the several informative features, including forward feature selection [13], sequential floating forward selection [14], minimal-redundancy-maximal-relevance algorithm [15], and l1-norm regularization [16]. In these methods, feature choices are mostly made regardless of existing relationships and structures between the features. However, considering them can be beneficial, potentially both for interpretability and to improve classification performance. To this end, structured sparsity-inducing norm regularization [17], which enforces structured group variable selection, was recently introduced and have been successfully applied in the fields of gene expression [18], text categorization [19], and speech assessment [20].

In this paper, we present a new feature selection and emotion classification method using a tree-structured sparse logistic regression model. Several acoustic features in spectral, prosodic, and glottal domains that are known as effective for emotion classification are first extracted. Then, hierarchical relationships of the acoustic features are defined as a tree-structure. Using this prior knowledge, we propose a structured sparse logistic regression model incorporated with the hierarchical structure of the features. The proposed model simultaneously and effectively addresses tree-structured sparse feature selection and emotion classification by introducing structured sparsity-inducing norm regularization. It can be expected that the feature may be selected on a structured group basis rather than an individual one. Therefore, we believe that more useful features can be selected to improve classification performance.

The remaining of the paper is organized as follows: the considered acoustic features are described in Section 2. In Section 3, we present the proposed tree-structured sparse logistic regression model for feature selection and emotion classification. Section 4 shows experimental results and finally our conclusions are summarized in Section 5.
2. Feature definition

In this section, acoustic features extracted from spectral, prosodic, and glottal domains are described and summarized in Table 1. Since the acoustic features are basically extracted with frame size of 30 ms and shift size of 15 ms, long-term statistics are calculated using mean (μ) and/or variance (σ²) in an utterance. The statistics are used as final acoustic features.

- Spectral features include MFCCs and the ratio of spectral flatness to spectral centroid (RSS) [21].
- Prosodic features include the fundamental frequency (F0) of the vocal fold vibration, the first three formant frequencies, and log energy. F0 and voicing frames are estimated using the summation of the residual harmonics (SRH) method [22], which is the state-of-the-art pitch tracking algorithm. The prosodic features are extracted for each voiced frame.
- Glottal source is estimated using iterative adaptive inverse filtering (IAIF) to perform the inverse filtering and recovering the glottal flow [23]. Then, normalized amplitude quotient (NAQ) [24], quasi-open quotient (QOQ) [25], H1-H2 ratio [26], harmonic richness factor (HRF) [27], and parabolic spectral parameter (PSP) [28] are extracted. Also, recently presented glottal features that are maxima dispersion quotient (MDQ) [29] and peak slope (PS) [30] are exploited. Finally, Teager energy operator (TEO) features [31] that use the area under the normalized Teager energy autocorrelation envelope are included. The glottal features are also extracted only for voiced frames.

3. Structured sparse logistic regression based feature selection and classification

3.1. Sparse logistic regression

The extracted features have meaningful information expressing the emotions in speech, but they may contain redundant information. In other words, not all extracted features may be effective to classify emotions. Therefore, choosing a proper feature set is very important in achieving better performance.

Let \( \mathbf{a} \in \mathbb{R}^p \) denote the \( p \)-dimensional speech features and \( y \in \{-1, +1\} \) be the associated (binary) class label. A logistic regression model is given by

\[
P(y|\mathbf{a}) = \frac{1}{1 + \exp(-y(\mathbf{\beta}^T \mathbf{a} + c))},
\]

where \( P(y|\mathbf{a}) \) is the conditional probability of the label \( y \), given the sample \( \mathbf{a}, \mathbf{\beta} \in \mathbb{R}^p \) is the \( p \)-dimensional model parameters, and \( c \in \mathbb{R} \) is the intercept (scalar). Suppose that we are given a set of \( m \) training data \( \{(\mathbf{a}_j, y_j)\}_{j=1}^m \), the likelihood function associated with these \( m \) samples is defined as \( \prod_{j=1}^m P(y_j|\mathbf{a}_j) \). The negative of the log-likelihood function is called the logistic loss, and the average logistic loss is defined as

\[
f(\mathbf{\beta}, c) = \frac{1}{m} \sum_{j=1}^m \log \left( 1 + \exp \left( y_j (\mathbf{\beta}^T \mathbf{a}_j + c) \right) \right).
\]

We can determine \( \mathbf{\beta} \) and \( c \) by minimizing the average logistic loss.

Sparse logistic regression models seek to predict an output by non-linearly combining a small subset of the features. To simultaneously address feature selection and prediction (classification), \( l_1 \)-norm regularization is utilized. The model parameters can be estimated by solving the following \( l_1 \)-norm regularized logistic regression problem [32]:

\[
\min_{\mathbf{\beta}, c} f(\mathbf{\beta}, c) + \lambda \|\mathbf{\beta}\|_1,
\]

where \( \|\cdot\|_1 \) denotes the \( l_1 \)-norm and \( \lambda > 0 \) is a regularization parameter. Note that the solution is found at the point that the level set of the empirical loss function in the first term of (3) is tangent to the surface of the parameter norm in the second term of (3). The \( l_2 \)-norm gives an isotropic round-shaped pattern that does not favor any specific direction in the parameter space while the \( l_1 \)-norm provides an anisotropic pattern, e.g., a diamond-shaped pattern in two dimensions, and exhibits some singular points due to the non-smoothness of \( l_1 \)-norm. Since these singular points are located along axis-aligned linear subspaces in \( \mathbb{R}^p \), if the level set of the empirical loss function is tangent to one of those points, sparse solutions are obtained, in which a small subset of parameters contains nonzero elements. In the sparse logistic model, it does not take into account any specific structures or possible relations among parameters since each parameter is treated independently. However, the acoustic features have hierarchical relations and therefore considering them during feature selection process may be significant in choosing a more appropriate feature set. To this end, we present the tree-structured sparse logistic regression in the next section.

3.2. Structured sparse logistic regression

The hierarchical structure of the acoustic features can be represented as a tree structure, e.g., spectral features can be divided into MFCCs and RSS, and the MFCCs include mean and variance features. For concise description, we define the tree structure as follows: for a tree \( Q \) with depth \( d \), let \( Q = \{G^i\}_i \),

<table>
<thead>
<tr>
<th>Acoustic features</th>
<th>Statistics</th>
<th>#Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MFCC (13 static + Δ + ΔΔ)</td>
<td>μ( ), σ²( )</td>
<td>78</td>
</tr>
<tr>
<td>RSS (1 static + Δ + ΔΔ)</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Prosodic features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F0 (1 static + Δ + ΔΔ)</td>
<td>μ( ), σ²( )</td>
<td>6</td>
</tr>
<tr>
<td>Formant (3 static + Δ + ΔΔ)</td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>Energy (1 static + Δ + ΔΔ)</td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Glottal features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAQ (1 static)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>QOQ (1 static)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>H1-H2 ratio (1 static)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>HRF (1 static)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>PSP (1 static)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>MDQ (1 static)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>PS (1 static)</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>TEO (16 static)</td>
<td></td>
<td>16</td>
</tr>
</tbody>
</table>
The nodes of depth 2 involve corresponding features belonging to their parent nodes, e.g., G1 includes MFCC and RSS, and G2 includes prosodic features such as F0, formants, and energy. The nodes of depth 2 involve corresponding features belonging to their parent nodes, e.g., G1 includes MFCC and RSS, and G2 includes RSS. The leaf nodes represent individual features associated with their parent nodes, e.g., G1 is the mean of first MFCC and G2 is the mean of second MFCC. Our tree structured sparse logistic model can then be formulated by the following tree-structured sparsity-inducing norm regularized logistic regression problem:

$$\min_{\beta,c} f(\beta,c) + \lambda \sum_{j=0}^{d} \sum_{k=1}^{n_j} w_j \| \beta_{G_j} \|_2,$$

where $\| \cdot \|_2$ denotes the $l_2$-norm, $w_j$ is the pre-defined weight for the group $G_j$, and $\beta_{G_j}$ is entries of parameters belonging to the group $G_j$. Note that the $l_1$-norm promotes sparsity while the $l_2$-norm encourages the grouping effect instead of sparsity. Therefore, regularizing the sum of the $l_2$-norms (i.e., the $l_2$ norm of $l_2$ norms) of a group of vectors in the second term of (4) induces the tree-structured group sparsity in the solution so that, only few child nodes contain nonzero elements if their parent node includes nonzero elements. That is, a variable or group (node) may be selected only if all its ancestors in a tree structure have already been selected, and if a group is not selected, then all its descendants are not also selected [17], [33]. Consequently, it is expected that the acoustic features are selected on a hierarchical group basis rather than an individual one. The group weights were determined proportional to the number of parameters belonging to the group as in [33].

### 4. Experiments

#### 4.1. Experimental setup

In order to evaluate the proposed method, we used Berlin database of emotional speech [37]. From the database, a total of 325 utterances for 4 emotions consisting of anger, happiness, neutral, and sadness spoken by 10 professional actors are considered in this experiments. To assess speaker independent performance, all experiments were performed in a leave-two-subjects-out cross validation. The classification accuracy was evaluated and averaged for all cross validation rounds. Also, sparsity was computed by measuring the ratio of the number of zero parameters and the number of all parameters at each cross validation round and then averaged. All prosodic and glottal features except TEO were extracted using COVAREP Toolkit [35] and SLEP Toolkit is utilized to train the structured sparse logistic regression (SSLR) [36]. All speech features were normalized to the interval between 0 and 1 using minimum-maximum normalization. To show the effectiveness of the proposed SSLR, the performance of SSLR is compared with those of conventional logistic regression (LR) and sparse logistic regression (SLR).

#### 4.2. Experimental results

Table 2 shows the performances of the proposed SSLR, SLR, and LR in terms of the sparsity and classification accuracy (CA). Here, the sparsity is set at the best performance of the CA. As can be seen, the performance of the SLR is better than with the LR, obtaining a CA of 76.2% using only 20% of all parameters. This result implies that not all features may be effective to classification and only a small set of features has an impact on classification task. The proposed SSLR obtained a CA of 80.3% and outperformed both SLR and LR performances, achieving 17.2% and 33.7% relative improvements in classification error reduction. For the SLR, the features are selected independently regardless of the

---

**Table 2. Performance evaluation of the proposed SSLR, SLR, and LR in terms of the sparsity and CA.**

<table>
<thead>
<tr>
<th>Model</th>
<th>Sparsity</th>
<th>CA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.00</td>
<td>70.3</td>
</tr>
<tr>
<td>SLR</td>
<td>0.80</td>
<td>76.2</td>
</tr>
<tr>
<td>SSLR</td>
<td>0.23</td>
<td>80.3</td>
</tr>
</tbody>
</table>

where $x \in \mathbb{R}^d$ denotes the feature vector for an unknown test utterance and $k$ is the emotion class index.
inherent structure of the features while for the SSLR, the feature choices are made by their hierarchical dependency. It ensures that considering the underlying structure of features allows choosing more appropriate feature set for the emotion classification.

To see how the classification performances were affected by the sparsity of the model parameters, the CA performances with varying sparsity for the SLR and SSLR are shown in Figure 3. Note that sparsity=0 corresponds to the baseline LR with full model parameters. As can be seen, the CA performance of the SLR produces a peak when sparsity is at 80% while the SSLR produce more stable and accurate results when sparsity is in the wide range from 20% to 50%. This result signifies that the proposed SSLR is less sensitive with varying sparsity. We also observe that although the sparsity is the same, we have different CA performances and different selected feature sets. Therefore, adopting an appropriate model is critical to performance improvements since the content of selected features is different depending on classification models.

Figure 4 presents the CA obtained for each emotion for the LR, SLR, and SSLR. It can be observed that the performances were improved with the SSLR for anger, neutral, and sadness while for happiness we obtained the worst performance with about 65% accuracy in all models. Looking into the results, we observed that happiness is mostly confused with anger, which is in line with the most studies [15]. This result suggests that the considered features are not suitable to capture the characteristics of the emotion of happiness, and that other features should be accounted.

Figure 5 indicates the distribution of the selected features for each emotion according to the spectral, prosodic, and glottal domains. Table 3 presents the ranking of the selected features. The ranking is determined based on the model parameter values. It can be observed that the spectral domain features largely have an impact on emotion classification for all emotions. For sadness, in particular, about 80% of selected features are spectral and all prosodic features are not selected. These results are in line with the ones obtained in [15] and [34]. Nevertheless, several glottal features are included in the best 5 features. From the analysis, we can conclude that combining acoustic features in different speech domain and selecting an appropriate feature set are very important in achieving better performance.

5. Conclusion

In this paper, a novel and effective method to automatically identify the speech emotions was proposed. First, we extracted possible acoustic features in three speech domains. To choose the best feature set and simultaneously to identify emotions, the tree-structured sparse logistic regression model incorporated with the hierarchical taxonomy of the acoustic features was proposed. The proposed method was evaluated on 4 emotions of Berlin emotional speech database in a speaker-independent manner. Experimental results showed that the proposed method provides significant improvement over the conventional logistic regression and sparse logistic regression, producing more reliable and stable results with varying sparsity. Also, our method gives the insight of selected features. Further works include the investigation to supplement more informative features and find optimal tree structures. Our framework could potentially be used for other paralinguistic applications of classification and assessment.

6. Acknowledgements

This work was supported by the Software Computing Technology Development Program, 14-824-09-012, Virtual Creatures with Digital DNA, funded by the Ministry of Science, ICT and Future Planning.
7. References


