Low-Dimensional Feature Space Derivation for Emotion Recognition

Jarosław Cichosz, Krzysztof Ślot

Institute of Electronics, Technical University of Lodz, Lodz, Poland
jarekcichosz@poczta.onet.pl; kslot@p.lodz.pl

Abstract
An objective of the paper was to determine a set of low-dimensional feature spaces that provide high emotion recognition rates. Candidates for target feature spaces were randomly drawn from a broad pool of speech signal parameters that comprised both commonly used characteristics and newly introduced features. As a result, several four-dimensional feature spaces that provide the highest emotion classification rates (68 %) on Polish language database, which we used in experiments, were identified.

1. Introduction
Emotion recognition becomes an increasingly important research direction in speech signal processing. This can be attributed to several reasons, such as human-computer interaction quality improvement, enhancement of speech and speaker recognition systems’ performance or, simply, an interest in assessment of speaker’s emotional state.

One of the key elements in pattern recognition is an appropriate selection of a feature space that is used in classification. Several speech signal characteristics have been proposed for the purpose of emotion representation so far, yet, none of the introduced parameters alone allows for obtaining satisfactory recognition rates. Therefore, a search for combinations of features, which provide good emotion discrimination as well as a search for new speech signal characteristics is still an open and important problem [1].

An objective of the reported research was to look for low-dimensional feature spaces that are appropriate for emotion recognition. A procedure that has been developed for a feature-space derivation attempts to determine subsets of a broad pool of speech signal characteristics, which are mutually uncorrelated and which provide high emotion recognition rates. An initial pool of features that we were using comprised both commonly used parameters as well as some newly introduced features, such as coefficients of linear and nonlinear regression of selected speech signal characteristics. As a result of the research, several candidates for four-dimensional feature spaces for emotion recognition were proposed and a set of features that appear most frequently across the winning feature spaces was identified.

Although the procedure used for feature space selection was executed on Polish language database, it is general and can be applied in case of any language.

The paper is organized in the following manner. Speech signal characteristics that are used in emotion recognition are outlined in Section 2. The procedure of the most discriminative feature-space derivation has been described in Section 3, while experiment results are summarized in Section 4.

2. Emotion descriptors
There exist numerous features of a speech signal that are considered in characterizing an emotional load of utterances. They can be grouped into a few categories, which reflect their physical properties. The commonly used parameters include pitch-derived measures (mean, deviations, range, global minima and maxima), signal energy descriptors, temporal characteristics (duration, vowel and pause duration) and others [2], [3], [4], [5], [6]. None of the above mentioned features, either used itself or used in combinations, proved to be sufficiently robust to ensure correct emotion classification. Therefore, we believe that there is still a need for refining emotion representation through a search for other characteristics of a speech signal. Taking this into account we decided to introduce several new candidates for speech signal characterization. The first group of features includes extensions to commonly used parameters, such as:

- Pitch evolution descriptors: mean of local maxima, mean of local minima and mean of local dynamic range
- Voiced speech characteristics: mean energy of voiced segments and a ratio of voiced fragment head and tail energies.

The second group of newly introduced parameters attempts to approximate temporal evolution of selected speech signal characteristics. Such an evolution often exhibits substantial differences for various emotional contents (Fig. 1). This issue has been addressed in [7], where we proposed to consider linear regression parameters \((a, b)\), derived for some speech characteristic \(s(t)\):

\[
\{a, b\} : \min \left\{ \mathbb{E}[(s(t) - (at + b))^2] \right\}
\]

as additional emotional speech features (Fig. 2a).

Emotion classification in feature spaces that include regression line slope yielded, in several cases, a significant increase in correct recognition rate [7].
A natural extension to modeling of speech signal evolution is to introduce also nonlinear regression. We decided to use third-order regression, i.e., to add another four coefficients:

$$\{A, B, C, D\} : E[(s(t) - (At^3 + Bt^2 + Ct + D))^2] \rightarrow \text{min} \tag{2}$$

to the pool of features considered in our experiments. One can observe (Fig. 1) that this might provide better approximation of long term trends for majority of analyzed sentences.

3. Feature space selection procedure

An objective of our research was to determine low-dimensional feature spaces that would provide correct emotion recognition. An initial feature set that was considered in our experiments was composed of \(N=49\) elements and included both commonly used characteristics (pitch, energy etc.) as well as the newly proposed parameters.

A procedure of a reduction of the original \(N\)-dimensional feature space is schematically depicted in Fig. 3. To derive a \(K\)-dimensional target feature space (where \(K<<N\)) each \(K\)-element subset of the considered \(N\)-element feature vector was a subject to the multi-step procedure.

![Figure 3: Feature space derivation procedure](image)

An objective of the first step was to ensure low correlation between elements of the currently considered \(K\)-element tuple. Therefore, mutual correlations were computed for all tuple elements. If any of these correlations exceeded some
Experimentally established threshold (which we set at the level of 0.3), a tuple was dropped from further analysis and the new $K$-element tuple was drawn from $N$ elements of the initial feature vector. Otherwise, the next step of the procedure - emotion recognition in the considered $K$-dimensional feature space - was executed. Linear Discriminant Analysis (LDA) [8], followed by the nearest-neighbor classification, was used as the recognition method. Due to relatively small size of available experimental data set, a "leave one sample out" strategy was adopted in the recognition procedure. All but one sample from each of six categories of emotions were used to train a classifier. Samples removed from the training set constituted test data. An outcome of the presented procedure has been executed on a database of thousands four-element tuples that can be drawn from of our original feature set, sorted according to correct emotion recognition rates.

4. Experiment results

The presented procedure has been executed on a database of Polish language sentences with an emotional load that falls into six different categories: anger, fear, sadness, boredom, joy and no emotion (a neutral emotion). The database contains 240 sentences uttered by four actors and four actresses. Each person was uttering the same five sentences, attempting to assign them with differing emotional load, thus producing six different sets of recordings.

To assess a quality of the prepared database material, the recordings have been evaluated by thirty subjects, through a procedure of classification of randomly generated samples. An average rate of correct recognition for this evaluation experiment was 72% (ranging from 60% to 84% for different subjects). This result is consistent with typically reported ratings, where average recognition rate spans from 60% to 80% [9].

A number of speaker-dependent emotion recognition experiments were executed, i.e. in each case a classifier was being designed and tested based on speech samples from the same speaker.

Four-dimensional feature space ($K=4$) has been selected as a target for the procedure that has been presented in the previous Section. We have chosen four-dimensional feature space as we think that there could be that many independent physical mechanisms involved in speech production, such as frequency, energy, temporal features and voice quality. Of over two hundred thousands four-element tuples that can be drawn from of our initial, 49-element feature space, 81% were dropped after the mutual correlation evaluation stage. Thus, a recognition procedure was executed in 18915 various feature spaces.

Recognition rates for the best $m$-tuples are summarized in Table 1. Abbreviations used for the features are the following:

- MP - mean pitch,
- ND – normalized duration,
- RLSP - regression line slope of pitch,
- %V – percentage of voiced speech,
- RLSI - regression line slope of intensity,
- %P - percentage pause in signal,
- MEDP – median pitch,
- MVIS – median intensity of voiced speech,
- MINIVS- minimum intensity of voiced speech,
- MEDI – median intensity,
- MINI – minimum intensity,
- STDI – standard deviation of intensity

One might consider the parameters given in Fig. 4 as good emotion descriptors. These parameters constitute potential elements of various feature spaces that can be spanned for emotion representation. One needs to bear in mind though that these features might be highly correlated. Correlation coefficients for winning parameters are shown in Table 2.

One can observe, that the set of parameters shown in Fig. 4 as good emotion descriptors. These parameters constitute potential elements of various feature spaces that can be spanned for emotion representation. One needs to bear in mind though that these features might be highly correlated. Correlation coefficients for winning parameters are shown in Table 2.

Limited size of the database makes results from Table 1 only a coarse attempt to the optimum feature space identification. To come up with more reliable evaluation of feature relevance in emotion recognition, we decided to make another analysis of the procedure results. Namely, a frequency of particular parameter appearance in sets of these $K$-tuples that produced high recognition rates (set arbitrarily at the level of 60%) were computed. Results of this analysis are shown in Fig. 4, where eight most frequent parameters are listed.
Table 2: Correlation coefficients for parameters shown in Fig. 4

<table>
<thead>
<tr>
<th></th>
<th>MP</th>
<th>ND</th>
<th>RLSp</th>
<th>%V</th>
<th>RLSI</th>
<th>%P</th>
<th>MEDP</th>
<th>MINP</th>
<th>MEDI</th>
<th>MINI</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean pitch (MP)</td>
<td><strong>1.00</strong></td>
<td>0.01</td>
<td>-0.22</td>
<td>0.19</td>
<td>0.27</td>
<td>-0.11</td>
<td>0.97</td>
<td>0.75</td>
<td>0.22</td>
<td>-0.14</td>
</tr>
<tr>
<td>normalized duration (ND)</td>
<td>0.01</td>
<td><strong>1.00</strong></td>
<td>0.09</td>
<td>-0.18</td>
<td><strong>-0.34</strong></td>
<td>0.24</td>
<td>0.00</td>
<td>0.10</td>
<td>-0.23</td>
<td>0.06</td>
</tr>
<tr>
<td>regression line slope of pitch (RLSP)</td>
<td>-0.22</td>
<td>0.09</td>
<td><strong>1.00</strong></td>
<td>-0.28</td>
<td>-0.17</td>
<td>0.21</td>
<td>-0.17</td>
<td>-0.02</td>
<td>-0.25</td>
<td>-0.06</td>
</tr>
<tr>
<td>percentage of voiced speech (%V)</td>
<td>0.19</td>
<td>-0.18</td>
<td>-0.28</td>
<td><strong>1.00</strong></td>
<td>0.32</td>
<td><strong>-0.46</strong></td>
<td>0.17</td>
<td>0.05</td>
<td>0.47</td>
<td>0.20</td>
</tr>
<tr>
<td>regression line slope of intensity (RLSI)</td>
<td>0.27</td>
<td>-0.34</td>
<td>-0.17</td>
<td><strong>0.32</strong></td>
<td><strong>1.00</strong></td>
<td>0.03</td>
<td>0.27</td>
<td>-0.05</td>
<td>0.15</td>
<td>-0.42</td>
</tr>
<tr>
<td>percentage pause in signal (%P)</td>
<td>-0.11</td>
<td>0.24</td>
<td>0.21</td>
<td><strong>-0.46</strong></td>
<td>0.03</td>
<td><strong>1.00</strong></td>
<td>-0.10</td>
<td>-0.15</td>
<td><strong>-0.76</strong></td>
<td>-0.68</td>
</tr>
<tr>
<td>median pitch (MEDP)</td>
<td>0.97</td>
<td>0.00</td>
<td>-0.17</td>
<td>0.17</td>
<td>0.27</td>
<td>-0.10</td>
<td><strong>1.00</strong></td>
<td>0.78</td>
<td>0.24</td>
<td>-0.14</td>
</tr>
<tr>
<td>minimum pitch (MINP)</td>
<td>0.75</td>
<td>0.10</td>
<td>-0.02</td>
<td>0.05</td>
<td>-0.05</td>
<td>-0.15</td>
<td>0.78</td>
<td><strong>1.00</strong></td>
<td>0.16</td>
<td>0.04</td>
</tr>
<tr>
<td>median intensity (MEDI)</td>
<td>0.22</td>
<td>-0.23</td>
<td>-0.25</td>
<td><strong>0.47</strong></td>
<td>0.15</td>
<td><strong>-0.76</strong></td>
<td>0.24</td>
<td>0.16</td>
<td><strong>1.00</strong></td>
<td>0.48</td>
</tr>
<tr>
<td>minimum intensity (MINI)</td>
<td>-0.14</td>
<td>0.06</td>
<td>-0.06</td>
<td>0.20</td>
<td><strong>-0.42</strong></td>
<td><strong>-0.68</strong></td>
<td>-0.14</td>
<td>0.04</td>
<td><strong>0.48</strong></td>
<td><strong>1.00</strong></td>
</tr>
</tbody>
</table>

attributed to the limited support of correct approximation of the relevant parameter. However, these coefficients can still appear superior if emotion recognition will be performed through an analysis of local speech-signal segments (which is the case when Hidden Markov Models are applied to model a speech).

5. Conclusion

A research on determination of a set of low-dimensional feature spaces that can be successfully applied for emotion recognition has been presented in the paper. The key element of the proposed feature-space selection procedure was to focus on mutual correlations between potential candidates. As a result, several feature spaces that provide recognition rates over 60%, for six-class, speaker-dependent recognition experiment, were selected. This result is better than the ones reported so far [6], which indicates that the proposed approach seems to be worth further studying.

We have also shown that modeling of temporal evolution of selected speech signal features can be a promising way of emotion representation. We are especially interested in modeling emotional speech with Hidden Markov Models with locally-computed nonlinear regression parameters as elements of observation vectors.

6. References