ON THE INFLUENCE OF QUANTIZATION ON THE IDENTIFIABILITY OF EMOTIONS FROM VOICE CODING PARAMETERS

Patrick Robitaille, Samuel Trempe, Philippe Gournay and Roch Lefebvre

Speech and Audio Research Group
Université de Sherbrooke
Sherbrooke (Québec) J1K 2R1 Canada

ABSTRACT

Although emotions play a major role in voice communication, the quality of their reproduction by low bit rate voice coders has never been investigated so far. This paper shows that the emotional state of a speaker can be identified automatically, with reasonable precision and accuracy, using conventional voice coding parameters (pitch, voicing, energy and LPC coefficients). It also shows that the performance of this identification degrades when these parameters are quantized, especially at lower rates (1200 bits/s). This suggests that quantization of speech parameters could be improved by targeting the faithful reproduction of important higher-level voice communication attributes such as emotions, rather than simply optimizing objective measures such as the signal-to-noise ratio, mean squared error and spectral distortion.

Index Terms— Speech coding, Vocoder, Parameters, Emotions, Identification

1. INTRODUCTION

Low bit rate voice coders, such as those used for satellite and military communications, rely on a compact yet mostly reversible parametric representation of the speech signal [1, 2]. By periodically transmitting pitch, voicing and energy values as well as vectors of LPC coefficients, excellent intelligibility, reasonable subjective quality, and decent speaker identifiability can be achieved at bit rates as low as 1200 bits/s [2].

Since the first developments in the early seventies [3], the impact of voice coding on intelligibility, subjective quality and communicability has been thoroughly studied [4]. The impact on speaker identifiability, i.e. the possibility for a listener to identify the speaker from its voice signal, has also been considered [5]. Nevertheless, despite its importance in speech communication, the impact of voice coding on the identifiability of emotions has so far been largely overlooked. Fear and joy, for example, have much in common in terms of prosody; they however bear a very different meaning. Being able to distinguish between these two emotions, conveyed by the voice of the speaker, represents a crucial piece of information for the listener. Furthermore, the emotional state of the speaker constitutes relevant contextual information that may increase the intelligibility of the communication [6].

In this study, automatic identification of emotions based on typical voice coding parameters is used to evaluate the impact of voice coding on the identifiability of emotions. Section 2 presents the feature extraction process, where primary parameters (pitch, voicing, energy and LPC coefficients) are being extracted from the speech signal, before being supplemented by secondary parameters (derivatives, ranges and statistics of the primary parameters). Section 3 deals with automatic identification of emotions, with an emphasis on the random forest approach which is used in this study. Section 4 presents detailed evaluation results obtained on various emotional speech databases, and with various degrees of quantization of the primary parameters. Finally, conclusions are drawn in the last section.

2. FEATURE EXTRACTION

The emotional state of a speaker can be determined from spectral and prosodic features extracted from his voice [7]. In essence, these features correspond to the parametric representation used in voice coders. Since the source code of the most recently standardized voice coder MELPe [2] is difficult to obtain and protected by copyrights, we used our own voice coder implementation called Harmonic-Stochastic eXcitation (HSX) [8]. The HSX obviously differs in some regards with other voice coders, but it is overall very representative of this family of low bit rate speech coders.

2.1 Primary parameters

The primary parameters used in this study are the pitch, voicing and energy values, along with sets of LPC coefficients. The HSX parameter extraction (also called analysis process) is represented in Figure 1. The speech signal is first downsampled to 8 kHz, then segmented into frames of 22.5 ms. We verified informally that the
downsampling operation does not alter too much the emotional content of the speech signal. One pitch lag, one voicing parameter (which, in the case of the HSX, consists in a variable cut-off frequency between a lower voiced band and an upper unvoiced band) and four energy values (computed once per subframe but on segments whose length depends on the pitch lag) are estimated for each frame. Two LPC filters are also estimated per speech frame, with the first one corresponding to the middle of the frame and the second one to the end of the frame. The order of these filters depends on the quantization rate (16 at 3200 bits/s, 12 at 2400 bits/s and 10 at 1200 bits/s). Further details about the HSX analysis process can be found in [8].

In equations 1 and 2, N is the number of voiced frames (i.e. speech frames for which the voicing cut-off frequency is greater than zero), while \( T_i \) and \( E_i \) are respectively the pitch lag and the energy.

In regard to the voicing parameter, its value is not used directly. For the purpose of our experiment, the voicing cut-off frequency is converted into a binary (voiced / unvoiced) information depending on whether it is zero (unvoiced) or greater than zero (voiced). We then compute the average length of voiced segments in the audio sample and the percentage of voiced frames in this sample.

Finally, LPC coefficients are not used directly, as that would entail an overwhelming number of features. Instead, we estimate the first and second formant frequencies by looking for the first two local maxima of the LPC spectral magnitude calculated using an FFT. Also, we estimate the spectral tilt (i.e. slope of the spectrum) by converting each set of LPC coefficients back to a set of correlation coefficients and calculating the ratio of the first two correlation coefficients \((-R_1/R_0)\).

This constitutes a total of 94 features (Table 1) that will be used for the automatic identification of the different emotions (section 3).

<table>
<thead>
<tr>
<th>Primary</th>
<th>Secondary</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch lag, Energy</td>
<td>First and second derivatives</td>
<td>Minimum, Maximum, Average, Range, Standard deviation, Kurtosis</td>
</tr>
<tr>
<td>Voicing</td>
<td>Average length of voice segments, Percentage of voiced frames</td>
<td></td>
</tr>
<tr>
<td>LPC coefficients</td>
<td>First and second formants, First and second derivative of those, Spectral slope coefficient</td>
<td>Minimum, Maximum, Average, Range, Standard deviation, Kurtosis</td>
</tr>
</tbody>
</table>

This constitutes a total of 94 features (Table 1) that will be used for the automatic identification of the different emotions (section 3).

**Fig. 1:** The HSX analysis process

With the corresponding HSX synthesis process, we again verified that this seemingly crude parametrization preserves the bulk of the emotional content of the speech signal.

### 2.2 Secondary parameters

The primary parameters are temporal values that cannot be fed directly to most automatic classifiers. Instead, in order to get a fixed number of features per audio sample, statistical values (minimum, maximum, range, average, standard deviation and kurtosis) are computed from the primary parameters and from secondary parameters that derive from these [9, 10, 11]. Using the minimum, maximum and range values may seem redundant but makes sense when using a random forest classifier.

Regarding the pitch lag and energy parameters, we use their first and second derivatives as secondary parameters. We also use the jitter and the shimmer [12]. The jitter characterizes the fluctuation of the pitch lag for all voiced frames within the audio sample:

\[
\frac{1}{N-1} \sum_{i=2}^{N} |T_i - T_{i-1}|, \quad (1)
\]

while the shimmer characterizes its energy fluctuation:

\[
\frac{1}{4N-1} \sum_{i=2}^{4N} 20 \log (E_i / E_{i-1}), \quad (2)
\]
quantization. The energy vectors are coded using differential, 4-dimensional vector quantization. The pairs of LPC filters are encoded jointly, in the line spectral frequency (LSF) domain, using predictive multi-stage split vector quantization.

At 1200 bits/s, three consecutive frames are grouped and encoded together as a super-frame. The quantizers were designed to optimize objective criteria such as mean squared error and spectral distortion. The details of this quantization process can be found in [8].

The bit allocation among primary parameters for each of the three quantized versions is summarized in Table 2.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>3200 bits/s</th>
<th>2400 bits/s</th>
<th>1200 bits/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch lag</td>
<td>7</td>
<td>6</td>
<td>6/3</td>
</tr>
<tr>
<td>Voicing</td>
<td>2</td>
<td>2</td>
<td>5/3</td>
</tr>
<tr>
<td>Energy</td>
<td>8</td>
<td>7</td>
<td>14/3</td>
</tr>
<tr>
<td>LPC Coeff.</td>
<td>53</td>
<td>37</td>
<td>55/3</td>
</tr>
<tr>
<td>Error Control and Synchron.</td>
<td>2</td>
<td>2</td>
<td>1/3</td>
</tr>
</tbody>
</table>

Table 2: Bit allocation for each of the three bit rates under consideration (in bits per frame of 22.5 ms)

3. AUTOMATIC IDENTIFICATION OF EMOTIONS USING RANDOM FORESTS

Many approaches have been proposed for automatic identification of emotions from a speech signal, including neural networks, Bayesian classifiers and support vector machines [13, 14]. The random forest approach seems adequate in a context where a large number of features are available, and has been shown to be successful for the classification of emotions [15].

A random forest is a collection of decision trees, each tree being based on a random subset of classification features. When an input vector of features needs to be classed, each decision tree is applied once, and the decision that gets the most votes constitute the final decision. Figure 2 shows an example of random forest with three trees. Further information about training and running a random forest classifier can be found in [16].

![Random Forest Diagram](image)

**Fig. 2:** An example of random forest with three decision trees. x is the input vector of features, k is the final decision.

4. EVALUATION RESULTS

This section presents detailed evaluation results obtained from various emotional speech databases, and with various quantization conditions for the primary parameters.

3.1 Emotional databases

There is currently a number of available emotional speech and audio-visual databases, featuring various emotions, languages (real or artificial), content types (sentences, dates and numbers), and recorded under various scenarios (acted or spontaneous) [13, 14]. For this study, we selected five easily accessible emotional databases [17-21], featuring at least four basic emotions (anger, sadness, joy and fear), and a fifth optional emotion (neutral). Some databases differentiate emotions that are very similar. For example, LDC distinguishes joy from pride, and hot anger from cold anger. We regrouped similar emotions into larger categories, such as “joy” and “anger”. We gathered a total of 2532 emotional speech samples. The distribution of these samples between databases and emotions is given in Table 3.

<table>
<thead>
<tr>
<th>Database</th>
<th>Language</th>
<th>Nb. of audio samples used</th>
</tr>
</thead>
</table>

Table 3: Distribution of the emotional audio samples in our combined database

3.2 Parameters of the random forest classifier

Two parameters need to be chosen when training a random forest classifier: the number of classification features in each tree, and the number of trees within the forest [22].

5952
Since complexity is not an issue in this study, we used a brute force approach to find the best combination of number of features (from 1 to 94) and number of trees (from 1 to 200) for each database and at each bit rate. Therefore, we trained 24 different random forest classifiers; one per database and per condition, each with its own number of classification features per tree and trees per forest. A typical forest contains 70 trees with 30 features per tree.

The classifier is implemented using the Statistics and Machine Learning toolbox from Matlab 2014. The function “treebagger” is used to train the random forest on a training set composed of an emotionally-representative, random selection of 80% of the database. The function “predict” is used to classify the test set composed by the remaining 20% of the database. The audio samples are randomized before training and testing.

### 3.3 Evaluation results

Figure 3 presents the classification rates across all emotions for the baseline and quantized conditions, for each of the five databases and for the combined database. The classification rate depends heavily on the database, and apparently very much on the number of audio samples in that database. The GVEESS database presents the highest classification rates, with 100% for all conditions except 1200 bits/s. This is due to the fact that the number of audio samples in this database is very small compared to the number of features and the number of decision trees. The classification rate is always above 60% for the baseline condition, as well as for the 3200 and 2400 bits/s conditions, which is satisfactory considering the context. There is very little degradation at 3200 and 2400 bits/s compared to the baseline condition, but the performance drops significantly at 1200 bits/s for most databases and for the combined database. This is likely due to the grouping of frames into superframes and the extremely coarse quantization of the primary parameters at this rate.

#### Table 4: Classification rates for each database and each bit rate (all emotions)

<table>
<thead>
<tr>
<th>Database</th>
<th>Baseline</th>
<th>3200 bits/s</th>
<th>2400 bits/s</th>
<th>1200 bits/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berlin</td>
<td>83.12%</td>
<td>85.71%</td>
<td>85.71%</td>
<td>72.73%</td>
</tr>
<tr>
<td>SAVEE</td>
<td>90.74%</td>
<td>88.89%</td>
<td>88.89%</td>
<td>70.37%</td>
</tr>
<tr>
<td>eNTERFACE</td>
<td>64.88%</td>
<td>64.29%</td>
<td>60.71%</td>
<td>52.98%</td>
</tr>
<tr>
<td>GVEESS</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>92.31%</td>
</tr>
<tr>
<td>LDC</td>
<td>65.82%</td>
<td>60.71%</td>
<td>60.71%</td>
<td>62.76%</td>
</tr>
<tr>
<td>Combined</td>
<td>63.08%</td>
<td>62.71%</td>
<td>62.52%</td>
<td>49.15%</td>
</tr>
</tbody>
</table>

#### Table 5: Confusion matrix for the combined emotional speech database at 1200 bits/s. Rows correspond to actual emotions and columns to predicted emotions.

<table>
<thead>
<tr>
<th></th>
<th>Anger</th>
<th>Sadness</th>
<th>Joy</th>
<th>Fear</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>49.18%</td>
<td>13.93%</td>
<td>18.03%</td>
<td>12.3%</td>
<td>6.56%</td>
</tr>
<tr>
<td>Sadness</td>
<td>11.54%</td>
<td>52.31%</td>
<td>14.62%</td>
<td>10%</td>
<td>11.54%</td>
</tr>
<tr>
<td>Joy</td>
<td>16.67%</td>
<td>8.33%</td>
<td>56.25%</td>
<td>16.66%</td>
<td>2.08%</td>
</tr>
<tr>
<td>Fear</td>
<td>17.76%</td>
<td>15.89%</td>
<td>24.3%</td>
<td>34.58%</td>
<td>7.48%</td>
</tr>
<tr>
<td>Neutral</td>
<td>6.58%</td>
<td>15.79%</td>
<td>14.47%</td>
<td>7.89%</td>
<td>55.26%</td>
</tr>
</tbody>
</table>

### 5. CONCLUSION

In this paper, we have shown that the emotional state of a speaker can be identified automatically, with reasonable precision and accuracy, using conventional voice coding parameters (pitch, voicing, energy and LPC coefficients). We have also shown that the performance of this identification degrades when the parameters are quantized, particularly at lower rates (1200 bits/s). These conclusions are strengthened by the fact that our experiences were conducted with a combination of five different databases (Berlin, SAVEE, eNTERFACE, GVEESS and LDC Emotional Prosody Speech and Transcripts).

Considering the importance of emotions in speech communication, these conclusions suggest that there is still room for improvement in low bit rate voice coding. Specifically, quantization of speech parameters could be improved by targeting the faithful reproduction of important higher-level voice communication attributes such as emotions, rather than simply optimizing objective measures, like signal-to-noise ratios, mean squared errors and spectral distortions.
6. REFERENCES


[22] Thais Mayumi Oshiro, Pedro Santoro Perez and José Augusto Baranauskas, “How many trees in a Random Forest?”, 8th International Conference on Machine Learning and Data mining in Pattern Recognition, Berlin, Germany, Volume 7376, pp. 154-168, June 2012