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

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

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].

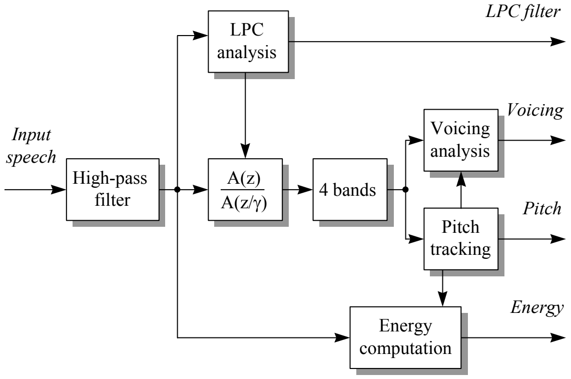

Fig. 1: The HSX analysis process

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:

$$\frac{1}{N-1}\sum_{i=2}^{N}\left|T_i - T_{i-1}\right|, \qquad (1)$$

while the shimmer characterizes its energy fluctuation:

$$\frac{1}{4N-1}\sum_{i=2}^{4N}\left|20\log(E_i/E_{i-1})\right|. \qquad (2)$$

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).

Primary	Secondary	Statistics
Pitch lag, Energy	First and second derivatives	Minimum, Maximum, Average, Range, Standard deviation, Kurtosis
	Jitter (pitch lag), Shimmer (energy)	
Voicing	Average length of voice segments, Percentage of voiced frames	
LPC coefficients	First and second formants, First and second derivative of those, Spectral slope coefficient	Minimum, Maximum, Average, Range, Standard deviation, Kurtosis

Table 1: List of primary and secondary parameters used for automatic identification of voice emotions

2.3 Quantization

As section 4 will present and compare experimental results obtained using a baseline version (voice analysis and synthesis without parameter quantization, the LPC order being equal to 16) and three quantized versions at decreasing bit rates (3200, 2400 and 1200 bits/s), here we shall describe the parameter quantization.

At 3200 and 2400 bits/s, the primary parameters are quantized on a frame by frame basis. The pitch lags and voicing cutoff frequencies are coded using absolute scalar quantization. The energy vectors are coded using differential, 4-dimentional 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.

	3200 bits/s	2400 bits/s	1200 bits/s
Pitch lag	7	6	6/3
Voicing	2	2	5/3
Energy	8	7	14/3
LPC Coeff.	53	37	55/3
Error Control and Synchron.	2	2	1/3

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].

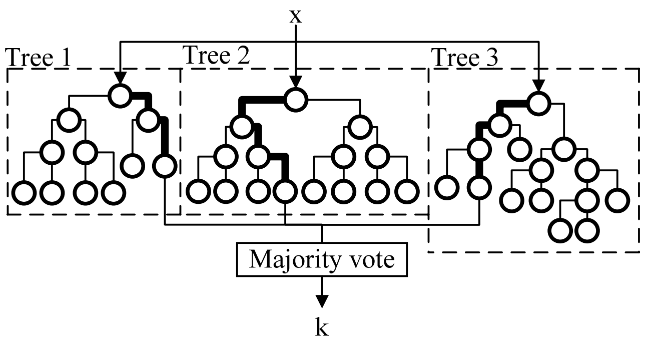

Fig. 2: An example of random forest with three decision trees. x is the input vector of features, k is the final decision.

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.

Database	**Language**	**Nb. of audio samples used**
Berlin [17]	German	Anger : 128 Sadness : 62 Joy:71 Fear :68 Neutral:54
SAVEE [18]	English	Anger :44 Sadness :45 Joy :45 Fear :44 Neutral : 90
eNTERFACE [19]	English	Anger :210 Sadness : 210 Joy : 207 Fear : 211
GVEESS [20]	Pseudo-German	Anger : 16 Sadness :16 Joy :16 Fear : 16
LDC Emotional Prosody Speech and Transcripts [21]	English	Anger :182 Sadness : 264 Joy : 223 Fear :167 Neutral :143

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].

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.

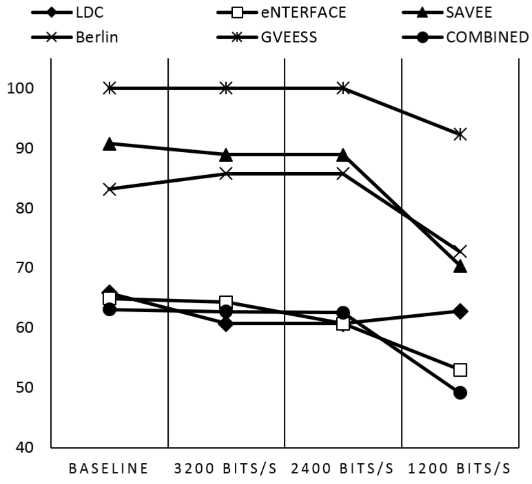

Fig. 3: Classification rates for each database and condition

	Baseline	3200 bits/s	2400 bits/s	1200 bits/s
Berlin	83.12 %	85.71 %	85.71 %	72.73 %
SAVEE	90.74 %	88.89 %	88.89 %	70.37 %
eNTERFACE	64.88 %	64.29 %	60.71 %	52.98 %
GVEESS	100 %	100 %	100 %	92.31 %
LDC	65.82 %	60.71 %	60.71 %	62.76 %
Combined	63.08 %	62.71 %	62.52 %	49.15 %

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

Table 4 presents the numerical values corresponding to Figure 3, whilst Table 5 shows an example of confusion matrix obtained at 1200 bits/s on the combined database. As stated above, confusing fear with joy is an easy (yet potentially serious) mistake, and mixing them up is the most frequent classification error.

	Anger	Sadness	Joy	Fear	Neutral
Anger	49.18 %	13.93 %	18.03 %	12.3 %	6.56 %
Sadness	11.54 %	52.31 %	14.62 %	10 %	11.54 %
Joy	16.67 %	8.33 %	56.25 %	16.66 %	2.08 %
Fear	17.76 %	15.89 %	24.3 %	34.58 %	7.48 %
Neutral	6.58 %	15.79 %	14.47 %	7.89 %	55.26 %

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

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

[1] Thomas E. Tremain, “The Government Standard Linear Predictive Coding Algorithm: LPC-10,” Speech Technology, pp. 40-49, April 1982

[2] Tian Wang, Kazuhito Koishida, Vladimir Cuperman, Allen Gersho and John S. Collura, “A 1200/2400 bps Coding Suite Based on MELP,” IEEE Workshop on Speech Coding, Tsukuba, Japan, October 6-9, 2002

[3] Bernard Gold, “A History of Vocoder Research at Lincoln Laboratory,” The Lincoln Laboratory Journal, Volume 3, Number 2, pp. 163-202, 1990

[4] Volodya Grancharov and W. Bastiaan Kleijn, “Speech Quality Assessment (chap. 5),” in Handbook of Speech Processing (pp. 83-100), A. Huang and M. Sondhi, Eds., New York, Springer, 2008

[5] Astrid Schmidt-Nielsen and Derek P. Brock, “Speaker Recognizability Testing For Voice Coders,” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Atlanta, Georgia, USA, 1996

[6] George A. Miller, George A. Heise and William Lichten, “The Intelligibility of Speech as a Function of the Context of Test Materials,” Journal of Experimental Psychology, Volume 41, Issue 5, pp. 329-335, May 1951

[7] K. Sreenivasa Rao and Shashidhar G. Koolagudi, *Robust Emotion Recognition using Spectral and Prosodic Features*, Springer, New York, 2013

[8] Philippe Gournay and Frédéric Chartier, “A 1200 bps HSX Speech Coder for Very Low Bit Rate Communications,” IEEE Workshop on Signal Processing Systems (SiPS'98), Boston, USA, October 8-10, 1998

[9] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi, “Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge,” Speech Communication, Volume 53, Issues 9–10, Pages 1062-1087, November–December 2011

[10] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou and Ioannis Giannoukos, “Features and Classifiers for Emotion Recognition from Speech: a Survey from 2000 to 2011,” Artificial Intelligence Review, Volume 43, Issue 2, pp 155-177, February 2015

[11] Brian D. Womack and John H. L. Hansen, “Classification of Speech under Stress using Target Driven Features,” Speech Communication, Volume 20, Issues 1-2, pp. 131-150, 1996

[12] Mireia Farrús and Javier Hernando, “Using Jitter and Shimmer in Speaker Verification,” IET Signal Processing, Volume 3, Issue 4, pp. 247-257, July 2009

[13] Dimitrios Ververidis and Constantine Kotropoulos, “Emotional Speech Recognition: Resources, Features, and Methods,” Speech Communication, Volume 48, Issue 9, pp. 1162-1181, September 2006

[14] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray, “Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases,” Pattern Recognition, Volume 44, Issue 3, Pages 572-587, March 2011

[15] Theodoros Iliou and Christos-Nikolaos Anagnostopoulos “Comparison of Different Classifiers for Emotion Recognition,” 13th Panhellenic Conference on Informatics, 10-12 Sept. 2009

[16] Leo Breiman, “Random Forests,” Machine Learning, Volume 45, Issue 1, Pages 5-32, October 2001

[17] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter Sendlmeier and Benjamin Weiss, “A Database of German Emotional Speech”, Interspeech'2005 - Eurospeech - 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005

[18] Sanaul Haq and Philip J. B. Jackson, “Speaker-Dependent Audio-Visual Emotion Recognition,” International Conference on Auditory-Visual Speech Processing, Norwich, UK, September 10-13, 2009

[19] Olivier Martin, Ioannis Kotsia, Benoît Macq, Ioannis Pitas, “The eNTERFACE' 05 Audio-Visual Emotion Database,” 22nd International Conference on Data Engineering, Atlanta, USA, 3-8 April 2006

[20] Rainer Banse and Klaus R. Scherer, “Acoustic Profiles in Vocal Emotion Expression,” Journal of Personality and Social Psychology, Volume 70, Issue 3, pp. 614-636, March 1996

[21] Mark Liberman, Kelly Davis, Murray Grossman, Nii Martey and John Bell, “Emotional Prosody Speech and Transcripts,” LDC2002S28, Web Download, Linguistic Data Consortium, Philadelphia, USA, 2002

[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