Recognizing low/high anger in speech for call centers

Article · January 2008

3 authors, including:

Fu-Ming Lee
Chaoyang University of Technology

Li Hua
Northwest A & F University

Recognizing Low/High Anger in Speech for Call Centers

FU-MING LEE*, LI-HUA LI, RU-YI HUANG
Department of Information Management
Chaoyang University of Technology
168 Jifong E. Rd., Wufong Township
Taichung County, 41349, Taiwan (R.O.C.)

Abstract: - Automatic multi-level anger recognition in speech is an important factor to enhance user satisfaction for call centers. In this research, three emotional states, namely, neutral, low anger, and high anger of acted corpora with telephone quality are specified for emotion recognition. The corpora are collected from amateur actors and, thereafter, verified by the actors themselves. The emotion recognizer is implemented by using Back-propagation Network (BPN) with acoustic features of speech examples. Due to the variation of expression methods by different person, the feature values of the training examples used are too complex to make the BPN model convergent. To overcome the problem, a codified method is developed to simplify the feature values. With the codified inputs, the BPN model and a comparative Decision Tree C5.0 give quite satisfactory test performances for anger recognition. Therefore, they can be used as a part of a decision support system for proper applications in call centers.

Key-Words: - Multi-level anger recognition, call center, acoustic features, Back-propagation Network

1 Introduction

Recently, there has been an increasing interest in automatic emotion recognition in speech [9][13][15][17][18], especially the negative emotion recognition. One motivation comes from the desire to develop human machine interfaces that are more adaptive and responsive to a user's behavior [5][9][10]. The other motivation comes from the need of a variety of applications for business [11][13][18]. In particular, in a call center environment, anger is identified as the negative and the most important emotion.

Three types of spoken material, namely, acted, induced, and real-life corpora, are generally collected for emotion analysis and recognition [2][7]. To test the performance of the emotion recognizer, we must know the actual emotional state of each test example of the corpora. Although, induced and real-life corpora are more natural than acted one, the state of each example of those corpora is unknown unless the speaker tells us the true answer right after the corpora being collected. In service practices, it is almost impossible to get the answer. Even if the state is labeled by several independent labelers, the inter-labeler agreement is very low [5]. Furthermore, even if all labelers achieve agreement, the answer may be wrong. On the contrary, the state of each example of acted corpora can be easily verified by the speaker. Therefore, in this research acted corpora are collected for building and evaluating the recognition model.

Features of the speech signal need to be extracted for automatic emotion recognition. There are large numbers of linguistic (lexical and dialogic) features and paralinguistic (acoustic, fluent, etc.) features that can be attributed to the emotional state of the signal [7]. Among them acoustic features are the most widely employed [9][13][19][20]. In this research linguistic features are excluded since the sentences spoken in acted corpora are controlled and expressed with the whole spectrum of considered emotions. Furthermore, because there is no agreement on the best set of relevant acoustic features for emotion recognition, our strategy is to use as many features as possible.

Many types of emotion recognizer of spoken material have been used, e.g. Artificial Neural Networks (ANN) [3][11][13][14][15][20], K-nearest Neighbors [6][9][10][11][13][15][16][20], Decision Trees [7][16], Gaussian models [5][12][15], and Support Vector Machines [7][11][15][16][18][20]. However, the most efficient model is still not well established and from published results appears to be data-dependent [18]. Because the major capability of ANN is their flexibility in approximating operation between inputs and outputs and one of the most popular models in ANN is Back-propagation

* Corresponding author (E-mail: fmlee@cyut.edu.tw; TEL: 886-4-23323000 ext. 4286)

Network (BPN), the recognition model is implemented by using BPN in this research. Furthermore, the performance of BPN is compared with the performance of Decision Tree C5.0, which is one of the most widely used recognizer.

The goal of this research is to create an emotion recognizer that can process telephone quality voice messages in real time and can be used as a part of a decision support system for proper applications in call centers. Three emotional states, namely, neutral, low anger, and high anger are considered for emotion recognition. This distinction is motivated by the wish to recognize low/high anger, especially low anger, which will be handled by the conciliation strategies. In our opinion multi-level anger recognition is an important factor to enhance caller satisfaction. However, recognition results are sparse in the literature. In the work of Burkhardt et al. [5], prosodic features of real-life corpora are extracted; an algorithm based on Gaussian densities are used to recognize emotions; the accuracy of the recognizer for no anger, low anger, and high anger emotions are 89%, 49% and 16%, respectively, which shows that the organizer does not distinguish between these emotions very well.

In this research, the acted corpora are collected from 48 amateur actors and, thereafter, verified by the actors themselves. From the corpora, 1401 examples are created and are partitioned into training set and test set, i.e. 1107 examples and 294 examples, respectively. Each data set has even examples for three emotional states. From each example, 47 acoustic features are extracted by using the Praat program [4] and, then, are normalized for developing and evaluating the emotion recognizers. A standard $F1$ measure is used to assess the performance of the emotion recognizers. Due to the variation of expression methods by different people, the BPN model can not converge in training stage. To overcome the convergent problem, a codified method is proposed to simplify the complexity of the normalized feature values. Thereafter, the BPN model with codified inputs is convergent; $F1$ measures of the model for test examples with neutral, low anger, high anger, and anger (i.e. low anger or high anger) emotional states are 71.5%, 47.0%, 64.4%, and 86.0%, respectively. $F1$ measures of Decision Tree C5.0 with codified inputs for test examples with neutral, low anger, high anger, and anger emotional states are 66.0%, 41.2%, 64.6%, and 82.4%, respectively. The results show that the performance of anger emotion detection is satisfactory. However, the performance of low anger emotion detection is poor since low anger is often confused with neutral or high anger.

2 Corpora Collection

In this research, three emotional states, namely, neutral, low anger, and high anger of acted corpora are specified for emotion analysis and recognition. The following is the details about the corpora and the collecting method.

- Recording equipments: Each utterance of the corpora is recorded by using a close-talk microphone with a recorder equipped in Microsoft Windows. To simulate the voice in telephone, the format of the recorded utterance is set at 8000 Hz, 8 bit, and mono.
- Amateur actors: There are 25 male actors and 23 female actors participating in this activity. They are undergraduate or graduate students with age between 20 and 30.
- Given sentences: Thirty short sentences are collected from the FAQs of websites of two telecommunication service providers in Taiwan. Each actor is asked to select ten of them which he (or she) supposes that it would be easier for him (or her) to speak at all three emotional states considered in this research.
- Recording environment: The recording activity takes place at a normal office. In a recording case, an actor speaks successively a sentence three times with neutral, low anger, and high anger emotional states, respectively.
- Quality verification: Right after a recording case the actor is asked to verify the quality of the utterance by playing it back. Unless the actor feels that the states of the utterance he portrays on demand are indeed consistent with the states he usually acts in the real life, the case will be recorded again.
- Emotional examples: Three emotional examples of an utterance are cut by a research member of us. After examining all the utterances, we find and discard 13 utterances with high background noises. Thus, the corpora of 1401 examples are created with 467 examples per emotional state.

3 Features Extraction

In this research, we intend to extract the acoustic features, namely, pitch, intensity, formant, pulse, shimmer, jitter, harmonics to noise ratio (HNR), and duration, by using the Praat program [4]. Before features extracting, every emotional example is filtered by a band-pass filter of the Praat program with band frequency set between 200 Hz and 3200 Hz which is close to the band frequency of voice in telephone. In features extracting, window length is set at 0.02 seconds. The formant frequencies are estimated by using the linear predicted coding (LPC) method. The prediction order is set as 16. Pitch floor and pitch ceiling are set according to the gender of the actor. For male's emotional example, the 2 parameters are set at 75 Hz and 300 Hz, respectively. For female's emotional example, the 2 parameters are set at 100 Hz and 500 Hz, respectively. The rest parameters in the Praat program are set as default values. The following is the details about the 47 features extracted from each emotional example.

- Pitch features: A pitch object represents periodicity candidates as a function of time, i.e. a pitch contour. Related pitch features are statistical properties of the pitch contour. Six pitch features are used in this research: maximum, minimum, mean, standard deviation, mean absolute slope, mean absolute slope without octave jumps.
- Intensity features: An intensity object represents an intensity contour at certain linearly spaced time points. Related intensity features are statistical properties of the intensity contour. Four intensity features are used in this research: maximum, minimum, mean, standard deviation.
- Formant features: A formant object represents spectral structure as a function of time, i.e. a formant contour. Related formant features are statistical properties of the formant contour. Twenty formant features are used in this research. The following is the details about the features.
  — First formant (F1): maximum, minimum, mean, standard deviation.
  — Second formant (F2): maximum, minimum, mean, standard deviation.
  — Third formant (F3): maximum, minimum, mean, standard deviation.
  — Fourth formant (F4): maximum, minimum, mean, standard deviation.
  — Bandwidths: B1, B2, B3, B4.
- Pulse features: It is generated at every point in the point process. Two related pulse features are used in this research: mean period, standard deviation period.
- Jitter features: Five jitter features are used in this research: jitter (local), jitter (local, absolute), jitter (rap), jitter (ppq5), jitter (ddp). The definitions of these features please referred to the Praat program [4].
- Shimmer feature: Five shimmer features are used in this research: shimmer (local), shimmer (apq3), shimmer (apq5), shimmer (apq11), shimmer (ddp). The definitions of these features please referred to the Praat program [4].
- HNR features: It represents the degree of acoustic periodicity. Four related HNR features are used in this research: maximum, minimum, mean, standard deviation.
- Duration feature: It represents the ratio of voiced frame's number and unvoiced frame's number.

4 Features Preprocessing

To train the emotion recognizer and test its performance, the 1401 emotional examples are partitioned into training set and test set. The training set consists of 1107 emotional examples created by randomly chosen 20 male actors and 18 female actors; the test set consists of 294 emotional examples created by the remaining 5 male actors and 5 female actors.

Because the units of the features are different and the ranges of the features' values of the emotional examples are diverse, all of the 47 features are normalized as follows. For all $e\in$training set$\cup$test set, let

$$\tilde{f}_i(e)=\frac{f_i(e)-f_{i,\min}}{f_{i,\max}-f_{i,\min}},$$
$$\bar{f}_i(e)=\begin{cases}1, & \text{if } \tilde{f}_i(e)>1,\\ \tilde{f}_i(e), & \text{if } 0\le\tilde{f}_i(e)\le 1,\\ 0, & \text{otherwise,}\end{cases} \tag{1}$$

where

$f_i(e)$ denotes the ith feature value of emotional example e,

$\bar{f}_i(e)$ denotes the normalized value of $f_i(e)$,

$f_{i,\max}$ denotes the 2nd maximum value of $\{f_i(e): e\in\text{training set}\}$,

$f_{i,\min}$ denotes the 2nd minimum value of $\{f_i(e): e\in\text{training set}\}$.

To avoid to get unusual extreme feature values, both $f_{i,\max}$ and $f_{i,\min}$ are determined as the 2nd extreme values mentioned above. Furthermore, to conform to practices, we suppose that the information about the test set is unknown in advance. Therefore, both $f_{i,\max}$ and $f_{i,\min}$ are determined only based on emotional examples of training set.

5 Performance Measure

In this research, $F1$ measure is used to assess the performance of recognizers. It has been used very commonly as a standard performance measure in the recognition problem, and is defined as follows.

$$F1=\frac{2PR}{P+R},\ R=\frac{n_c}{m_c},\ P=\frac{n_c}{m'_c} \tag{2}$$

where n_c denotes the number of test examples whose emotional state is correctly recognized as state c, m_c denotes the number of test examples with emotional state c, and m'_c denotes the number of test examples whose emotional state is recognized as emotion state c.

6 BPN with Normalized Inputs

The BPN model proposed in this research is depicted on Fig. 1. It consists of three kinds of layers, i.e. input layer, hidden layer, and output layer. The inputs consists of 47 normalized features. The number of hidden nodes is set as 25. The number of output nodes is set as the number of emotional states. In training stage, the target values of the training example for the outputs are set according to the emotional state of the example described in Table 1. In test stage, the recognized emotional state of the test example is determined according to the output with highest value. For example, if $y_l(e)$ has the highest value, the emotional state of the test example is recognized as low anger.

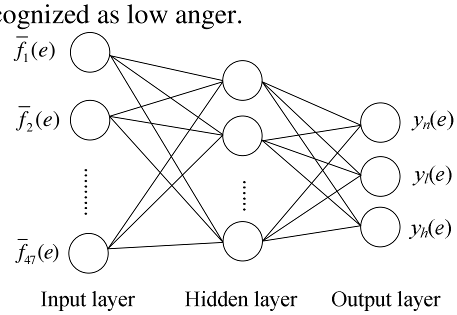

Fig. 1 The BPN model

In the experiments, all the 1107 training examples are used to train the BPN model by using the Matlab's Neural Network Toolbox. However, the BPN model can not converge. To find out the reason, the training set and the test set are regrouped by discarding examples with certain emotional state; the 47 features are preprocessed as (1) with regrouped training set and test set; the BPN model is reformed by discarding the output node corresponding to the certain emotional state; the target values of the training example for the remaining 2 outputs are set as shown in Table 1, and the reformed BPN model is trained by using the regrouped training examples. The reformed BPN model can not converge either in the case where high anger emotional examples are discarded or in the case where neutral emotional examples are discarded. Nevertheless, in the case where low anger emotional examples are discarded, the reformed BPN model is convergent and its performance is shown in Table 2. The $F1$ measures for test examples with neutral or high anger emotional states are above 85%.

Table 1 Target value of training example for outputs

output / emotional state	$y_n(e)$	$y_l(e)$	$y_h(e)$
neutral	1	0	0
low anger	0	1	0
high anger	0	0	1

Table 2 Performance of BPN without node $y_l(e)$

recognized as / examples	neutral	high anger	measures (%)		
			P	R	$F1$
neutral	83	15	86.5	84.7	85.6
high anger	13	85	85.0	86.7	85.8

In collecting the acted corpora, we observe that the ways to speak a given sentence with neutral, low anger, and high anger emotion are quite different for different person. Thus, for each feature, the feature values of the three emotional examples and the patterns of the values' distribution are quite different for various persons. Furthermore, for the three emotional states, some actors' expression methods are hard to discriminate, especially the methods between neutral and low anger or between low anger and high anger. This is the reason why the BPN model and the reformed BPN models with node $y_l(e)$ can not converge, and the reformed BPN model without node $y_l(e)$ are convergent and has excellent test performance.

Because the quality of the examples is verified by the actor of the examples right after a recording case, the low anger examples are still valid and should be considered. To overcome the convergent problem of the BPN model, the complexity of the normalized feature values of the emotional examples is simplified by codifying method described in the next section.

7 Features Codification

To codify ith normalized feature value of any emotional example, ith normalized feature values of the 1107 training examples are collected. They are ordered by their values and are partitioned into five segments depicted on Fig. 2, in which four partition points are determined as (3).

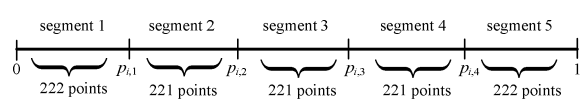

Fig. 2 Partition points for the ith normalized feature

$$p_{i,s} = \frac{\bar{f}_{i,s,lp} + \bar{f}_{i,s+1,fp}}{2}, \text{ for } i = 1, 2, ..., 47,\ s = 1, 2, 3, 4, \text{ and} \quad (3)$$

for the ith normalized feature

$p_{i,s}$ denotes the sth partition point,

$\bar{f}_{i,s,lp}$ denotes the value of the last point at segment s,

$\bar{f}_{i,s+1,fp}$ denotes the value of the first point at segment s+1.

Every normalized feature value of an emotional example is codified into four digits defined in Table 3. For instance, when *i*th normalized feature value of an emotional example belongs in certain segment depicted on Fig. 2, the value is codified into four binary digits corresponding to the segment. In Table 3, the binary codes are designed such that the hamming distance of any two codes represents the segment distance of the corresponding two segments. For example, the hamming distance of two codes corresponding to segment 1 and segment 5 is 4. The details are described as in Table 4.

Table 3 The binary code for each segment

segment	seg. 1	seg. 2	seg. 3	seg. 4	seg. 5
binary code	1000	1100	1110	1111	0111

Table 4 The hamming distance of any two codes

segment	seg. 1	seg. 2	seg. 3	seg. 4	seg. 5
seg. 1	0	1	2	3	4
seg. 2	1	0	1	2	3
seg. 3	2	1	0	1	2
seg. 4	3	2	1	0	1
seg. 5	4	3	2	1	0

8 BPN with Codified Inputs

To accommodate the codified inputs, the BPN model is redesigned with 188 binary input nodes and 95 hidden nodes. The number of outputs and the target value of the training example are designed by the same way for the BPN model mentioned in Section 6. The BPN model with codified inputs is trained by the 1107 training examples. Finally, the BPN model with codified inputs is converged and the performance of the model is shown in Table 5. The $F1$ measures for test examples with neutral, low anger, and high anger emotional states are 71.5%, 47.0%, and 64.4%, respectively.

Table 5 Performance of BPN with codified inputs

recognized as / examples	neutral	low anger	high anger	measures (%)		
				P	R	$F1$
neutral	69	18	11	72.6	70.4	71.5
low anger	22	44	32	49.4	44.9	47.0
high anger	4	27	67	60.9	68.4	64.4

Table 6 Performance of BPN with codified inputs for neutral and anger emotions

recognized as / examples	neutral	anger	measures (%)		
			P	R	$F1$
neutral	69	29	72.6	70.4	71.5
anger	26	170	85.4	86.7	86.0

As mentioned in Section 6, low anger emotional examples are often confused with neutral or high anger emotional examples. Thus, although the convergent problem of the BPN model can be overcome by codifying the complex normalized features, the performance of the BPN model with codified inputs are still not good for the three emotional states, especially low anger. However, with respect to applications in call center, the performance for anger, i.e. low anger or high anger, emotion recognition is the main concern. Therefore, after combining the results of both low anger and high anger, the performance of the BPN model with codified inputs is recalculated as Table 6. The $F1$ measure for test examples with anger emotional state is 86.0%.

To compare the performance of the BPN model with codified inputs, Decision Tree C5.0 with codified inputs is developed and evaluated with the same training set and the test set by using SPSS Clementine. The performance of Decision Tree C5.0 is shown in Table 7 and Table 8. The $F1$ measures for test examples with neutral, low anger, high anger, and anger emotional states are 66.0%, 41.2%, 64.6%, and 82.4%, respectively. The results of the two models show that the performance of anger emotion detection is satisfactory. However, the performance of low anger emotion detection is poor since low anger is often confused with neutral or high anger. From the experiments, we notice that the BPN model is superior to the Decision Tree C5.0.

Table 7 Performance of Decision Tree C5.0 with codified inputs

recognized as / examples	neutral	low anger	high anger	measures (%)		
				P	R	$F1$
neutral	66	17	15	64.7	67.3	66.0
low anger	26	34	38	50.7	34.7	41.2
high anger	10	16	72	57.6	73.5	64.6

Table 8 Performance of Decision Tree C5.0 with codified inputs for neutral and anger emotions

recognized as / examples	neutral	anger	measures (%)		
			P	R	$F1$
neutral	66	32	64.7	67.3	66.0
anger	36	160	83.3	81.6	82.4

9 Conclusions

Automatic multi-level anger recognition in speech is an important factor to enhance user satisfaction for call centers. To recognize the level of anger, one difficulty comes from the lack of clear definition or description to distinct the levels of anger and the other difficulty comes from the need of actual emotional state information of the telephone speech examples. In our opinion, the caller himself is the only person who exactly knows the answer and the emotional state of the caller is the main concern of call centers.

In this research, neutral, low anger, and high anger emotions are considered. To have the exact emotional state information of speech examples, the acted corpora are verified by the actor himself. Due to the variation of expression methods by different people, the feature values of the training examples used are too complex to make the BPN model convergent. To overcome the problem, a codified method is developed to simplify the feature values. With the codified inputs, the results of the BPN model and a comparative Decision Tree C5.0 show that the performance of anger (i.e. low anger or high anger) emotion detection is satisfactory. Therefore, they can be used as a part of a decision support system for proper applications in call centers.

References:

[1] J. Adell, A. Bonafonte, and D. Escudero, “Analysis of Prosodic Features: Towards Modeling of Emotional and Pragmatic Attributes of Speech,” *XXI Cogreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 2005.

[2] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, “Desperately Seeking Emotions or: Actors, Wizards, and Human Beings,” *Proc. of The Inter. Speech Communication Association Workshop on Speech and Emotion*, 2000, pp. 195-200.

[3] M. W. Bhatti, Y. Wang, and L. Guan, “A Neural Network Approach for Human Emotion Recognition in Speech,” *Proc. of IEEE Inter. Symposium on Circuits and Systems*, Vol. 2, 2004, pp. 23-26.

[4] P. Boersma and D. Weenink, “Praat: Doing Phonetics by Computer (Version 4.4.33),” http://www.praat.org/, 2006.

[5] F. Burkhardt, M. V. Ballegooy, R. Englert, and R. Huber, “An Emotion-Aware Voice Portal,” *Proc. of The 16th Conf. on Electronic Speech Signal Processing*, 2005, pp. 123-131.

[6] F. Dellaert, T. Polzin, and A. Waibel, “Recognizing Emotion in Speech,” *Proc. of The Inter. Conf. on Spoken Language Processing*, 1996, pp. 1970-1973.

[7] L. Devillers, L. Vidrascu, and L. Lamel, “Challenges in Real-Life Emotion Annotation and Machine Learning Based Detection,” *Neural Networks*, Vol. 18, No. 4, 2005, pp. 407-422.

[8] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, “Emotion Recognition by Speech Signals,” *Proc. of The 8th European Conf. on Speech Communication and Technology*, 2003, pp. 125-128.

[9] C. M. Lee, S. S. Narayanan, and R. Pieraccini, “Recognition of Negative Emotions from the Speech Signal,” *Proc. of Automatic Speech Recognition and Understanding*, 2001, pp. 240-243.

[10] C. M. Lee and S. S. Narayanan, “Toward Detecting Emotions in Spoken Dialogs,” *IEEE Trans. on Speech and Audio Processing*, Vol. 13, No. 2, 2005, pp. 293-303.

[11] D. Morrison, R. Wang, and L. C. De Silva, “Ensemble Methods for Spoken Emotion Recognition in Call-Centers,” *Speech Communication*, Vol. 49, 2007, pp. 98-112.

[12] D. Neiberg, K. Elenius, and K. Laskowski, “Emotion Recognition in Spontaneous Speech Using GMMs,” *Proc. of The 19th Inter. Conf. on Spoken Language Processing*, 2006, pp. 809-812.

[13] V. A. Petrushin, “Emotion Recognition in Speech Signal: Experimental Study, Development, and Application,” *Proc. of The 6th Inter. Conf. on Spoken Language Processing*, 2000, pp. 951-954.

[14] H. Sato, Y. Mitsukura, M. Fukumi, and N. Akamatsu, “Emotional Speech Classification with Prosodic Parameters by Using Neural Networks,” *Proc. of The 7th Australian and New Zealand Intelligent Information Systems Conf.*, 2001, pp. 18-22.

[15] B. Schuller, G. Rigoll, and M. Lang, “Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine - Belief Network Architecture,” *Proc. of Internat Conf. on Acoustics, Speech, and Signal Processing*, Vol. 1, 2004, pp. 577-580.

[16] M. Shami and W. Verhelst, “An Evaluation of the Robustness of Existing Supervised Machine Learning Approaches to The Classification of Emotions in Speech,” *Speech Communication*, Vol. 49, 2007, pp. 201-212.

[17] D. Ververidis and C. Kotropoulos, “Emotional Speech Recognition: Resources, Features, and Methods,” *Speech Communication*, Vol. 48, 2006, pp. 1162-1181.

[18] L. Vidrascu and L. Devillers, “Detection of Real-Life Emotions in Call Centers,” *Proc. of The 18th Inter. Conf. on Spoken Language Processing*, 2005, pp. 1841-1844.

[19] S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, “An Acoustic Study of Emotions Expressed in Speech,” *Proc. of Internat Conf. on Spoken Language Processing*, Vol. 1, 2004, pp. 2193-2196.

[20] F. Yu, E. Chang, Y.-Q. Xu, and H.-Y. Shum, “Emotion Detection from Speech to Enrich Multimedia Content,” *Proc. of The 2nd IEEE Pacific Rim Conf. on Multimedia*, 2001, pp. 550-557.

ISSN: 1790-5117

ISBN: 978-960-6766-44-2