Acoustic Noise Classification Using Selective Discrete Wavelet Transform-Based Mel-Frequency Cepstral Coefficient

Salinna Abdullah, Majid Zamani and Andreas Demosthenous
Department of Electronic and Electrical Engineering, University College London (UCL),
Torrington Place, London WC1E 7JE, UK
E-mail: salinna.abdullah.13@ucl.ac.uk, m.zamani@ucl.ac.uk; a.demosthenous@ucl.ac.uk

***Abstract*— A feature extraction method through wavelet Mel-Frequency Cepstral Coefficients (MFCCs) is proposed for acoustic noise classification. The method combined with a wavelet sub-band selection technique and a feedforward neural network with two hidden layers, is a promising solution for a compact acoustic noise classification system that could be added to speech enhancement systems and deployed in hearing devices such as cochlear implants. The technique leads to higher classification accuracies (with a mean of 95.25%) across three SNR values, a significantly smaller feature set with 16 features, a reduced memory requirement, faster training convergence and lower computation cost by a factor of 0.69 in comparison to the traditional Short-Time Fourier Transform-based (STFT-based) technique.**

Keywords—Acoustic noise classification, neural network, dimensionality reduction, mel-frequency cepstral coefficients, discrete wavelet transform, sub-band selection.

I. Introduction

Noise suppression and speech enhancement algorithms employed in cochlear implants, mobile communication and automatic speech recognition have been shown to perform well in noisy conditions to a certain extent. Specifically, in the field of cochlear implants, conventional speech enhancement algorithms that utilise spectral subtraction [1], subspace projection [2] and statistical-model [3] algorithms generally achieve significant improvement of speech intelligibility in stationary noise, but modest or non-significant improvement for speech intelligibility in non-stationary noise. The success of these algorithms has been limited partly because although they have been created to accommodate all acoustic environments, they only show optimal speech enhancement in a limited range of background, usually stationary, noise scenarios [1]. Real-world auditory environments are challenging in that they encompass a large variety of temporal and spectral characteristics that require a more adaptable approach to speech enhancement. Therefore, there is great interest in developing noise suppression/speech enhancement algorithms that are much more adaptable to the acoustic environment, which suggests the need for an acoustic noise classification algorithm to be embedded into speech enhancement techniques.

In this paper, a novel methodology for achieving a compact and robust acoustic noise classification system is introduced. The method is based on the extraction of wavelet parameters from the original signal for classification using the Discrete Wavelet Transform (DWT). From the Hurst exponents and ℓ_2 -norm (used as an energy function) extracted from the wavelet channels, more suitable DWT channels are selected for the extraction of a subset of Mel-Frequency Cepstral Coefficients (MFCCs) to be added to the MFCC feature vector generated from the original signal. The classification process involves a neural network, namely a feedforward Multi-Layer Perceptron (MLP). The decision to extract features from wavelet decomposition was motivated by the finding that feature extraction in the joint time-frequency domain is more suited for effective representation of non-stationary characteristics (e.g. trends, discontinuities and repeating patterns) of audio signals in comparison to feature extraction in the time or spectral domain [4]. Furthermore, DWT exhibits a multi-resolution approach by analysing different frequencies with different resolutions, in contrast to the Short-Time Fourier Transform (STFT), which uses a fixed window size for all frequencies.

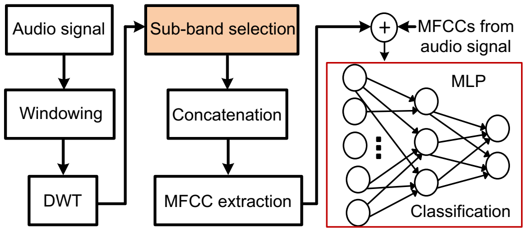

Fig. 1. The overall process of the proposed acoustic noise classification algorithm. Sub-band selection is a component proposed in this paper to more effectively select features for classification training and testing.

Similar approaches to the proposed algorithm have been explored by researchers for other purposes. [5] explored wavelet-based mel-scaled features for acoustic scene classifications. They reported that their proposed wavelet-based system, when combined with a support vector machine classifier, performed considerably better when compared with two benchmark systems, one based on MFCCs and Gaussian mixture models, and another based on log mel-band energies and MLP. In [6], wavelet features extracted from suitable preprocessed electroencephalogram channels are used for human emotion recognition with promising outcomes. This paper encompasses certain aspects from both [5] and [6], since mel-scaled features are extracted from suitable channels in the wavelet domain for acoustic noise classification.

The simulation results obtained from using clean speech utterances mixed with different types of noise at different

This work was supported by an EPSRC Industrial Strategy Studentship to S. Abdullah.

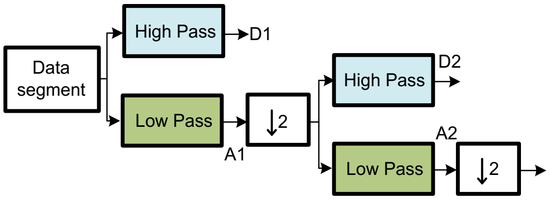

Fig. 2. Two-level wavelet tree decomposition where the framed signal is fed into a high pass and low pass filter to yield the detail and approximation coefficients respectively. The approximation coefficients are then downsampled by two before the decomposition is repeated again to give the coefficients at the subsequent level.

SNRs show that the proposed approach is capable of achieving high recognition rates, provide more discriminative features for training and smaller memory requirement with a small set of 16 features. A compact and robust acoustic noise classification system such as the one proposed here, could easily be added to the front-end of many acoustics processing-based algorithms and could potentially be implemented in hearing devices such as hearing aids and cochlear implants, where a small, low-powered and robust system is desired.

The rest of the paper is organised as follows. In Section II, the essential components of the proposed noise classification system are described. Section III describes the methodology for the experiments conducted which include the datasets and evaluation metrics used. Section IV discusses the observations and results of the comparisons made. Finally, the subsequent section concludes the paper.

II. Proposed Wavelet Cepstrum Feature for Compact Acoustic Noise Classification

In this section, the broad overview of the proposed acoustic noise classification system is first presented and then the important components combined to make up the proposed system are each explained in individual sub-sections.

A. Overall Concept

The main steps of the proposed approach are summarized in the flowchart given in Fig. 1. The feature extraction and classification are done on a frame-by-frame basis, where the continuous audio signal is blocked in frames of 35 ms using a Hamming window. To avoid a loss of information, a frame overlap of 10 ms is used. The Hurst exponent and the ℓ_2-norm are used as selection criteria to determine the three most suitable wavelet decomposed sub-bands to be used for the MFCC feature extraction. The MFCC is compact, discriminative and is commonly used in speech processing applications including speaker identification and sound recognition. Finally, an MLP architecture is used for the classification process.

B. Discrete Wavelet Transform

In a Fourier transform a signal is represented through a linear combination of indefinitely long sine waves that are not localised in time. In contrast, the DWT expands a signal into a set of basis functions known as wavelets. These wavelets are localised in time and the convolution of a signal with them provides the frequency information of the signal accompanied by its time information. Wavelet transform offers high frequency resolution at low frequencies and high time resolution at high frequencies.

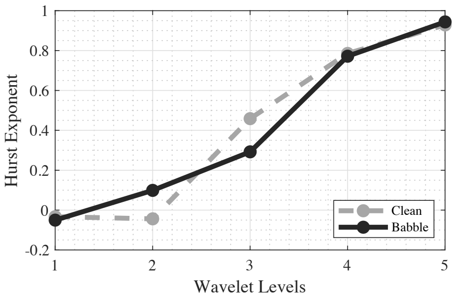

Fig. 3. Example H values (Hurst exponents) calculated for up to 5 detail wavelet levels for a random 35-ms frame of clean speech and the same speech corrupted by babble noise at -5 dB.

When DWT is applied to a signal, it is represented as a series of approximations where the low pass version of the decomposition corresponds to the coarse approximations and the high pass version corresponds to the detail information. This decomposition is achieved through a filter bank structure as shown in Fig. 2, where higher level approximation coefficients are passed through a highpass and a lowpass filter, and then downsampled by two to compute both the detail and approximation coefficients at a lower level. This tree structure is repeated for a multi-level decomposition. In this paper, the Daubechies wavelet of order 4 (db4) is used as the basis wavelet and the number of decomposition levels chosen is five.

C. Hurst Exponent

The Hurst exponent, H, was proposed in [7] to compose a speech feature vector and was successfully applied to speaker recognition. It is a statistical measure used to classify time series and is related to its spectral characteristics. $H = 0.5$ indicates a random series (e.g. white noise); $0 < H < 0.5$ indicates an anti-persistent series, often associated with high frequencies; and $0.5 < H < 1$ indicates a persistent, trend reinforcing series where low frequencies are prominent. Therefore, a larger value of H is a more regular and less erratic process than a smaller one. Fig. 3 shows sample values of H at different detail levels extracted from a random frame of clean and noisy speech. The grey dashed line represents the H values estimated from a TIMIT (a speech corpus recorded at Texas Instruments and transcribed at Massachusetts Institute of Technology [8]) clean speech signal and the solid line represents the same speech signal corrupted by babble noise obtained from the NOISEX [9] dataset at -5 dB, which is explained further in Section III.A. In this particular example, a larger variation of H values between the clean and corrupted signal could be found in detail wavelet levels 2 and 3. It was found that larger variation of H values is often found for detail levels exhibiting H values at or below 0.5, suggesting that the noise classes used are better distinguished at higher frequencies.

D. ℓ_2-Norm

The ℓ_2-norm can be used as the energy function to identify the dominant or less dominant frequency sub-bands obtained from the wavelet decomposition. It is calculated as shown in Equation 1 [10]:

$$\|\vec{x}\|_2 = \sqrt{\sum_{i=1}^{n} (x_i)^2} \qquad (1)$$

where $\vec{x}$ is a vector and $\|\vec{x}\|_2$ denotes the ℓ_2-norm of the vector. It was found empirically that for better classification of the noise classes used in the experiment, wavelet decomposition detail levels that are less dominated by speech should be used for feature extraction. This suggests that selection of detail sub-bands with lower ℓ_2-norm are more suitable for training and testing since sub-bands with lower energy tend to possess fewer speech components. It was also found that the selection of one approximation sub-band with two detail sub-bands led to better classification performance. In contrast, the approximation sub-band that possesses the most energy is selected for feature extraction leading to better classification performance since this selection criterion often results in a sub-band that encompasses a wide range of frequencies, thus ensuring that to a certain extent the proposed method also takes into account the information in other frequencies (i.e., not just high frequency information) in the classification process. Therefore, a sub-band is shortlisted for selection when it exhibits H values at or below 0.5 and subsequently, three most suitable sub-bands according to the ℓ_2-norm criterion will proceed to undergo the feature extraction process.

E. Mel-frequency Cepstral Coefficient

In the calculation of MFCC features, the Fourier transform is taken from a framed signal and the magnitude of the resulting spectrum is warped by the mel-scale. In contrast to the normal cepstrum which uses linearly spaced frequency bands, the mel-scale frequency bands are approximately linearly spaced below 1 kHz and logarithmic above. Such a frequency band configuration makes them more closely resemble the human auditory system response. Equation 2 is used to convert f hertz into m Mel [11]:

$$m = 2595 \log_{10}\left(\frac{f}{700} + 1\right). \qquad (2)$$

The log of the mel-scale warped spectrum is subsequently obtained before a discrete cosine transform is applied to obtain the cepstral coefficients. In this paper, a 32-channel mel-scale is used to obtain 13 MFCCs of the audio signal. For the MFCCs of the concatenated wavelet sub-bands, again 13 MFCCs are extracted but only a subset (i.e., 3) of the 13 coefficients are kept for the feature vector. The chosen subset is the first three MFCCs since this subset and subset number were empirically (i.e., the classification performance was observed each time the number of coefficients was incremented by 1) found to provide better classification performance without enlarging the feature vector significantly. The selection of the first three coefficients is also supported by [12], which in their investigation for finding the MFCCs that can better discriminate between vowels reports that the first three coefficients achieved top Fisher scores, indicating that the first three MFCCs are more discriminative than the rest.

F. Multi-layer Peceptron

The MLP is perhaps the most popular model in neural networks. It consists of an input layer, one or more hidden layers and an output layer with feedforward connections in between the layers. Each layer consists of nodes and, with the exception of the input nodes, each node is a neuron that uses an activation function that defines the output of the node for a given input or a set of inputs. In this case, the symmetric sigmoid transfer ('tansig') function is used as the activation function. An MLP is trained with a backpropagation algorithm. In this process the network weights and biases are adjusted using a training algorithm to minimise the prediction error, measured by a cost (loss) function, between the predicted output and the desired output. The MLP used is a 4-layer perceptron neuron network (including 1 linear input and output layer) consisting of 2 hidden layers with 10 neurons each. It uses the Mean Square Error (MSE) as a cost function and is trained with the Levenberg-Marquardt [13, 14] learning algorithm. The training is stopped when the magnitude of gradient used to adjust the network weights and biases is less than $1e^{-5}$ or when the maximum training epoch of 1000 is reached.

III. Methodology

In this section the experimental setup, datasets, evaluation metrics and other feature extraction methods employed for assessing the proposed system are introduced.

A. Datasets

The method for generating the training and testing datasets was adapted from that in [15]. 1000 randomly chosen utterances from the TIMIT training set were used as the training utterances and 100 utterances from the TIMIT core test set, consisting of 192 utterances from unseen speakers of both genders, were used as the test utterances. For the training and testing noises, 5 noises from the NOISEX dataset were used. The noises are a mix of 4-minute long stationary and nonstationary noises that include babble noise, factory noise, pink noise, Volvo (car) noise and white noise. A sampling frequency of 16 kHz was used throughout the experiment. In order to create the training sets and to avoid using the same noise segments for both training and testing, random cuts of the first 2 minutes of each noise were used to mix with the training utterances at -5 and 0 dB SNR. The test mixtures were in turn a mix of random cuts of the last 2 minutes of each noise and the test utterances at -5, 0 and 5 dB SNR. 5 dB SNR is an unseen condition.

B. Evaluation Methods

For classification, the rate of correct classifications made by the trained model (CAcc), is used.

The computation cost is measured in terms of the number of additions (or subtractions) and the weighted number of multiplications (or divisions) needed to execute the entire algorithm, which depends upon the implementation of the individual sub-algorithms contained within it. For instance, the computation cost of the DWT, MFCC extraction, sub-band selection and MLP are combined to give the approximate overall cost for executing the proposed

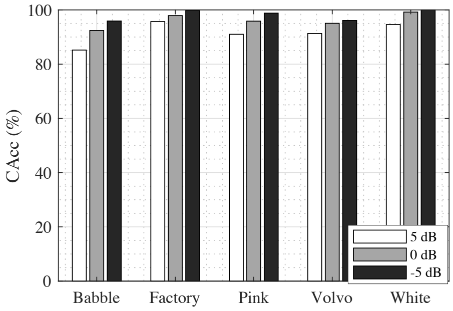

Fig. 4. Classification accuracy obtained by the proposed method for TIMIT utterances contaminated by NOIZEUS babble, factory, pink, Volvo and white noise at 5, 0, and -5 dB.

TABLE I. CLASSIFICATION ACCURACY COMPARISON

Feature extraction method	Feature vector Size	Average CAcc (%)			
		SNR			Mean
		5 dB	0 dB	-5 dB	
STFT	281	91.12	95.72	95.78	94.21
MRCG + Δ + ΔΔ	384	98.16	99.52	99.84	99.17
Signal MFCCs	13	90.64	94.20	96.38	93.74
Signal + all decomposition levels MFCCs	16	91.22	95.84	97.86	94.97
Signal + selected decomposition levels MFCCs	16	91.56	96.08	98.12	95.25

framework. The cost of implementing the initial windowing process is disregarded since this is a cost shared by all the algorithms compared in this study.

The approximate time taken to train an epoch is also evaluated to note the time needed for the neural network to reach convergence. The results obtained from the proposed algorithm are compared with those obtained when the following feature extraction methods are used: (1) STFT; (2) multi-resolution cochleagram (MRCG) [16]; (3) 13 MFCCs obtained from the audio; and (4) a subset of MFCCs obtained from all decomposition levels (i.e. absent of sub-band selection) used in combination with MFCCs from the audio. The MRCG, proposed by [16], reports good classification performance. It is a combination of four cochleagrams that encode power distributions of an audio signal, also in the time-frequency representation at different resolutions. The high-resolution cochleagram captures the local information while the three low-resolution cochleagrams capture more global spectrotemporal contexts at different scales. The addition of delta (Δ) and double delta (ΔΔ) to yield the MRCG + Δ + ΔΔ feature set was suggested by [16] to better capture temporal dynamics of the signal. The MRCG + Δ + ΔΔ feature set resulted in a dimensionality of 384 (32×4×3) for each 35-ms frame.

IV. RESULTS AND DISCUSSION

Table I shows the average CAcc achieved by the different feature extraction methods described in Section III.B for three different SNR (5, 0 and -5 dB) values. The proposed method which involves sub-band selection has a higher mean classification accuracy of 95.25% than STFT (94.85%), MFCC feature set of the signal only (93.74%), or MFCC feature set of the signal concatenated with the MFCC subset extracted from all wavelet decomposition levels (94.97%) is used for feature extraction. The increased classification accuracy obtained from adding the sub-band selection capability shows that the selection of more informative wavelet sub-bands for feature extraction can indeed lead to improved classification performance. The higher classification performance exhibited by the wavelet-based methods validates that a multi-resolution approach by analysing different frequencies with different resolutions is beneficial in the task of acoustic noise classification. The classification accuracy obtained with MRCG + Δ + ΔΔ is the highest for every SNR scenario tested. However, this comes with a much higher computation cost, and its feature size (384 features) for a single frame is larger than that provided by the proposed method by a factor of 24. Such a large feature size, when fed into the same neural network configuration means having more nodes at the input layer and thus, more parameters to take into consideration during training and testing. In addition to a wider network structure, this leads to a much longer training time and larger hardware memory. The observation that most of the feature extraction methods assessed perform better when tested with -5 dB SNR than when tested with 0 dB SNR demonstrates their ability to learn from the noise components rather than the speech. Results obtained for the 5 dB SNR are the poorest, but still provided CAcc at above 90%. They show that the listed feature extraction methods are capable of generalising to unseen SNR conditions. Fig. 4 shows a more granular view of the classification performance results obtained from the proposed method, where the classification performance for each noise class at different SNR is plotted as a bar in a bar chart. The babble noise (relatively more non-stationary) is the most challenging noise to recognise in the classification task whilst the white noise is the easiest.

TABLE II. COMPUTATION COST AND TRAINING TIME COMPARISON

Feature	Computation cost*	Train time**
STFT	187 137	19
MRCG + Δ + ΔΔ	2 697 476	36
Signal MFCCs	53 357	< 1
Signal + all decomposition levels MFCCs	61 350	< 1
Signal + selected decomposition levels MFCCs	129 332	< 1

* Cost per 35-ms frame based on ComputComp = $N_{add(sub)} + 10N_{mult(div)}$, where $N_{add(sub)}$ is the number of additions (or subtractions), and $N_{mult(div)}$ is the number of multiplications (or divisions) [17].

** Approximate time taken in seconds to train one epoch in MLP with 10,000 samples.

The estimated computation costs and training time per epoch of each of the feature extraction methods are listed in Table II. The STFT-based classification method is the second most efficient to compute but requires a larger memory due to its feature size and much longer training time. The proposed method has reduced the number of arithmetic operations needed to compute the MLP and the overall computation cost by a factor of around 0.69 relative to the STFT-based method. A mean improvement of 1.04% in the classification accuracy, reduced memory and significantly less training time have been achieved. The proposed method is effective in reducing the training time and boosting the robustness of the classification framework with a smaller memory requirement. This outcome demonstrates the importance of well thought out feature extraction procedure and invites more research into achieving a good compromise between high classification accuracy and low computation cost in acoustic noise classification systems.

This paper explores the use of Hurst exponents and ℓ_2-norm for wavelet sub-band selection, and wavelet-based MFCCs for classification. Although this approach has shown promising results, more extensive research is needed to identify complementary sub-band selection and feature extraction schemes with even lower computation demands. While the proposed method shows good classification accuracy for trained noise types, its generalisation performance on unseen noise types has not been assessed. It is important that the acoustic classification system is robust even when tested with new noise types. Future work will include assessing the generalisation performance of the proposed method and improving its computation cost and CAcc in higher SNR conditions.

V. Conclusion

In this paper, a method for achieving a compact classification system with high classification accuracy in speech noise classification has been introduced. For achieving a compact feature vector size, a set of selection criteria has been formulated for an efficient selection of the wavelet sub-bands for feature extraction as well as for the selection of more discriminative subset of the MFCCs. A simple feedforward multi-layer perceptron has been used for the recognition process. Results show that recognition rates are improved by a mean of 1.51% with the introduction of three additional wavelet-level features and are higher in the cases of low SNR. These further come with a decrease in computation cost by a factor of around 0.69 when compared to a conventional STFT-based classification method. Overall, the low-cost wavelet and DNN-based framework is promising for implementation in compact systems such as hearing devices.

References

[1] L. Yang and Q. Fu, “Spectral subtraction-based speech enhancement for cochlear implant patients in background noise,” The Journal of the Acoustical Society of America, vol. 117, no. 3, pp. 1001-1004, Mar. 2005.

[2] P. Loizou, A. Lobo and Y. Hu, “Subspace algorithms for noise reduction in cochlear implants,” The Journal of the Acoustical Society of America, vol. 118, no. 5, pp. 2791-2793, Nov. 2005.

[3] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” IEEE Transactions Audio, Speech, Language Processing, vol. 32, no. 6, pp. 1109-1121, Dec. 1984.

[4] B. Ghoraani and S. Krishnan, “Time–frequency matrix feature extraction and classification of environmental audio signals,” IEEE Transactions Audio, Speech, Language Processing, vol. 19, no. 7, pp. 2197-2209, Sept. 2011.

[5] S. Waldekar and G. Saha, “Wavelet Transform Based Mel-scaled Features for Acoustic Scene Classification”, Interspeech 2018, Sept. 2018.

[6] M. Islam and M. Ahmad, “Wavelet Analysis Based Classification of Emotion from EEG Signal”, 2019 International Conference on Electrical, Computer and Communication Engineering, Feb. 2019.

[7] R. Sant’ Ana, R. Coelho, and A. Alcaim, “Text-independent speaker recognition based on the hurst parameter and the multidimensional fractional brownian motion model,” IEEE Transactions Audio, Speech, Language Processing, vol. 14, no. 3, pp. 931-940, May 2006.

[8] J. Garofolo, “DARPA TIMIT acoustic-phonetic continous speech corpus,” Gaithersburg, MD, USA: National Institute of Standards Technology, 1993.

[9] A. Varga and H. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” Speech Communication, vol. 12, pp. 247-251, 1993.

[10] T. Chang, C.-C. J. Kuo, “Texture analysis and classification with tree-structured wavelet transform”, IEEE Transactions on Image Processing, vol. 2, pp. 429-441, Oct. 1993.

[11] M. A. Hossan, S. Memon and M. A. Gregory, “A novel approach for MFCC feature extraction,” 2010 4th International Conference on Signal Processing and Communication Systems, pp. 1-5, Jan. 2011.

[12] S. Hegde, K. K. Achary and S. Shetty, “Feature selection using Fisher’s ratio technique for automatic speech recognition,” International Journal on Cybernetics and Informatics, vol. 4, no. 2, pp. 45-51, Apr. 2015.

[13] K. Levenberg, “A method for the solution of certain non-linear problems in least squares”, Quarterly of Applied Mathematics, vol. 2, no. 2, pp. 164-168, Jul. 1944.

[14] D. W. Marquardt, “An algorithm for the least-squares estimation of nonlinear parameters,” SIAM Journal of Applied Mathematics, vol. 11, no. 2, pp. 431-441, Jun. 1963.

[15] Y. Wang, A. Narayanan and D. Wang, “On training targets for supervised speech separation,” IEEE/ACM Transactions Audio, Speech, Language Processing, vol. 22, no. 12, pp. 1849-1858, Dec. 2014.

[16] J. Chen, Y. Wang and D. Wang, “A Feature Study for Classification-Based Speech Separation at Low Signal-to-Noise Ratios”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 12, pp. 1993-2002, Dec. 2014.

[17] M. Zamani and A. Demosthenous, “Feature extraction using extrema sampling of discrete derivatives for spike sorting in implantable upper limb neural prostheses”, IEEE Transactions on Neural Systems and Rehabilitation. Eng., vol. 22, no. 4, pp. 716–726, Jul. 2014.