Connectionist Temporal Classification-based Sound Event Encoder for Converting Sound Events into Onomatopoeic Representations

Koichi Miyazaki[1], Tomoki Hayashi[1], Tomoki Toda[2], Kazuya Takeda[1]
[1]Graduate School of Information Science, Nagoya University
[2]Information Technology Center, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464–8601, Japan
{miyazaki.koichi, hayashi.tomoki}@g.sp.m.is.nagoya-u.ac.jp, tomoki@itcs.nagoya-u.ac.jp, kazuya.takeda@nagoya-u.ac.jp

***Abstract*—In this paper, we propose a sound event encoder for converting sound events into their onomatopoeic representations. The proposed method uses connectionist temporal classification (CTC) as an end-to-end approach to directly convert a sequence of feature vectors of each sound event into a corresponding onomatopoeic word representation which accurately represents each sound and can be intuitively understood. Moreover, to address the issue of the ambiguity of onomatopoeic representations among different individuals, we develop a database of sound events and their corresponding typical onomatopoeic representations as accepted by multiple listeners. To evaluate the performance of our proposed method, we conduct objective and subjective evaluations. Experimental results demonstrate that the proposed sound event encoder is capable of converting sound events into their onomatopoeic representations with a 74.5% subjective acceptability rating, and that use of typical onomatopoeic representations, as approved by multiple subjects, yields significant improvement, resulting in an acceptability rate of 81.8%.**

***Index Terms*—connectionist temporal classification, sound event, onomatopoeia, sound transcription**

I. Introduction

Acoustic wave caused by a specific physical event can be referred to as a sound event. Many studies related to sound events have been conducted in recent years, such as the development of sound event detection and classification techniques [1] which use of sound events to understand sound environments. These techniques have great potential for use in practical applications, such as multimedia retrieval [2], sound environment analysis [3], and monitoring systems [4]. On the other hand, since sound events are nonverbal, they do not always have symbolic representations that can be used to different them, in the way a written language is used to represent speech sounds. Therefore, it can be difficult to represent various kinds of sound events in a unified framework. It would be useful if we could develop a consistent and universal symbolic representation system for arbitrary sound events.

In this study, we focus on onomatopoeic representation as a possible system for symbolic representation of sound events. Onomatopoeic representation is based on human perception of sound events and imitation of their sounds using lexical phrases, e.g., the description of the sound of a firecracker or gunshot using the word "bang" in English. Although the phonology of the onomatopoeic representation correlates to acoustic sounds, it can have many variations resulting from differences in the perceptions of individuals or cultural influences background. For example, in Japan "wan wan" is used to represent the sound of a dog's bark, while people in English-speaking countries will usually use "bow wow" or "woof woof." Commonly used onomatopoeic representations are transmitted together with culture, and therefore they often differ over different regions, but they usually have some characteristics in common, such as common vowels, consonant types or number of syllables [5]. Onomatopoeic representation is intuitively understandable, making it possible for us to represent even new or unknown sounds like voice imitation [6]. It is often used as an effective expression technique in comics [7]. A method of searching for sound effects or music using onomatopoeic representations as a search query has been proposed [8] [9]. It is expected that the development of a technique to convert arbitrary sound events into their corresponding onomatopoeic representations could be useful in various existing applications, and could also have great potential to lead to new applications.

In this paper, we propose an end-to-end sound event encoder to convert arbitrary sound event signals into their corresponding onomatopoeic representations by employing Connectionist Temporal Classification (CTC) to directly convert a sequence of acoustic feature vectors extracted from a given sound event signal into a phoneme sequence corresponding to an onomatopoeic representation. Moreover, to address the issue of the ambiguity of onomatopoeic representations among different individuals, we develop a database consisting of various sound events and their corresponding onomatopoeic representations as accepted by multiple listeners. A subjective evaluation is then conducted to investigate the effectiveness of the proposed sound event encoder and demonstrate that 1) it is capable of converting sound events into acceptable onomatopoeic representations, 2) by having the onomatopoeic representations in our database vetted by multiple subjects, we are able to achieve a significant improvement in performance.

ISBN 978-90-827970-1-5

II. RELATED WORK

There are two main approaches for converting sound events into their corresponding onomatopoeic representations, 1) a classification approach, and 2) an automatic speech recognition (ASR) approach. In the classification approach, possible onomatopoeic representations need to be defined in advance [10]. A given sound event signal is then converted into one of the possible representations using a classifier. As a result, it is difficult to handle undefined sound events using a classification approach. The ASR approach is more flexible. As an example of a typical method based on this approach, Ishihara et al. proposed a system for the conversion of a sound event into an onomatopoeic representation in Japanese [11]. Assuming that one Japanese syllable explicitly corresponds to a single acoustic waveform segment, a given waveform is first segmented using various acoustic methods. Each waveform segment is then classified into a possible Japanese syllable using a Gaussian mixture model and a hidden Markov model (GMM-HMM) with mel frequency cepstrum coefficients (MFCC). Although this method can handle arbitrary sound events, it is difficult to properly segment the waveform signals, and final conversion performance is severely affected by segmentation accuracy since this is a sequential process.

III. DATABASE CONSTRUCTION

A. Onomatopoeia database

To realize our onomatopoeia conversion system, we first needed to construct an onomatopoeia database. We used 9,720 sound event samples selected from the RWCP sound scene database [12], which contained 100 kinds of sounds such as the sound of hitting a piece of wood with a mallet, the sound of an electronic toy, a person whistling, and so on. Each sound sample was converted into a 16-bit, 16 kHz monaural signal. One adult, Japanese male listened to these sound samples and manually transcribed each of them using the following annotation rules:

- Transcribe each sound into Japanese katakana symbols,
- Follow a standard form of Japanese onomatopoeic representation, e.g., use standard Japanese syllables, where each syllable consists of a vowel or a pair of a consonant and a vowel, do not start a geminate consonant or a syllabic nasal (N), and so on,
- Use a geminate consonant at the end of crunching sounds,
- Use long vowels at the end of a sustaining sound,
- Do not use long vowels consecutively,
- For a repeating sound, repeat the corresponding Japanese syllables,
- Do not try to represent changes in pitch.

As a result of this process, we obtained 492 unique onomatopoeic representations.

B. Database of typical onomatopoeic representations

Because suitable onomatopoeic representations depend on individual interpretations, representations assigned by one person may not be accepted by others. However, Oishi et al. [5] reported that typical onomatopoeic representations which are widely accepted do exist. In order to obtain more widely accepted onomatopoeic representations, we conducted a double-check procedure using 10 male and female subjects as follows:

1. Divide the onomatopoeia database into 10 subsets,
2. Assign one subset to each subject,
3. Have subjects listen to each sound sample and judge whether the assigned onomatopoeic representation is acceptable,
4. If it is acceptable, move on to the next sample,
5. If it is not acceptable, assign a more suitable onomatopoeic representation to the sample, then move on to the next sample,
6. Repeat steps 3-5 to evaluate all of the samples in the subset,
7. Define the acceptable onomatopoeic representations as “typical” onomatopoeic representations for the corresponding sound samples,
8. For sound samples which were determined to have had unacceptable original onomatopoeic representations, conduct the entire process from step 1 using only the newly assign onomatopoeic representations and a new subject.

Note that a different subset was assigned to each subject for the subsequent evaluations. After repeating this procedure three times, we obtained onomatopoeic representations for all of the sound samples which were accepted by at least two of the subjects. The resulting database of typical onomatopoeic representations for 695 unique sounds was then used as the database for our proposed system.

IV. CTC-BASED SOUND EVENT ENCODER SYSTEM

A. System overview

An overview of our proposed method, separated into training and test phases, is shown in Fig. 1. In the training phase, the sound event signal is divided into $40\,\mathrm{ms}$ windows with 50 % overlap to calculate a 40-dimensional log mel filter bank feature. The statistics of the extracted features are calculated over training data to perform normalization, making mean and variance of each dimension of the features 0 and 1, respectively. Then a bidirectional long short-term memory recurrent neural network (BLSTM) [13], [14] with projection layers [15], [16] (Fig. 2) is trained using the normalized features and the onomatopoeia labels on the basis of the objective function of connectionist temporal classification (CTC) process. In the test phase, as in the training phase, features are calculated from the sound event signal and normalized using the statistics of the training data. Finally, best path decoding of CTC using the normalized features is performed, resulting in the estimated onomatopoeic representations.

B. Connectionist Temporal Classification

CTC is a framework which can handle differences in the lengths of input and output sequences. As a result, it can be applied to our unsegmented sequences. In CTC, blank symbols (“_”) are added to the set of output symbols to allow

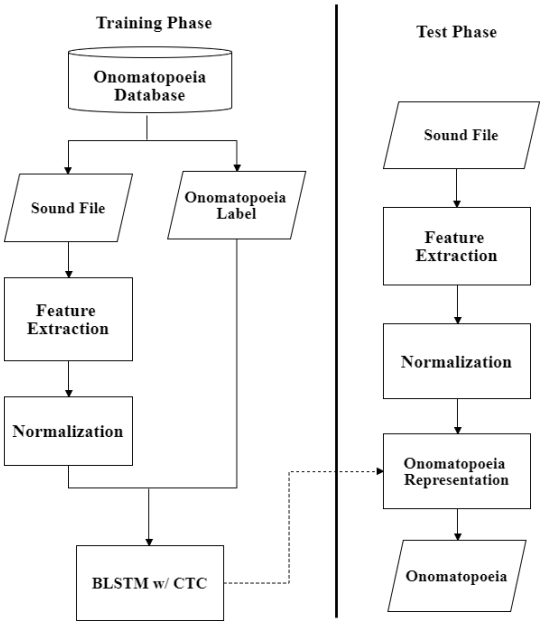

Fig. 1. Overview of proposed method

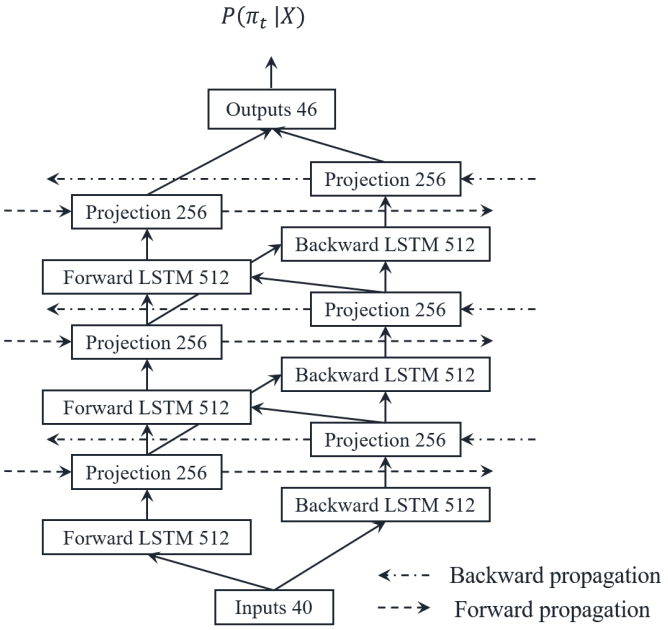

Fig. 2. Structure of BLSTM with projection layers

an output sequence to use a redundant description to adjust its length to equal that of the input sequence. For example, redundant descriptions with a length of three corresponding to the sequence (a b) are (_ a b), (a a b), (a b b), (a _ b), and (a b_). Thus, there are many redundant descriptions of the output sequence π. We define a mapping function $\beta(\pi \rightarrow l)$ to remove redundancy by deleting consecutive symbols and blank symbols from the redundant descriptions.

The posterior probability of the redundant output sequence $\pi = [\pi_1, \pi_2, \cdots, \pi_T]$ for the input sequence $X = [x_1, x_2, \cdots, x_T]$ is calculated using the following equation:

$$p(\pi|X) = \prod_{t=1}^{T} p(\pi_t|X), \tag{1}$$

where $p(\pi_t|X)$ is the posterior probability of the symbol π_t at time t, which is modeled by the BLSTM as follows:

$$p(\pi_t|X) = \mathrm{BLSTM}_t(X). \tag{2}$$

The posterior probability of output sequence l is obtained as the sum of the output probabilities for all of the redundant output sequences as follows:

$$p(l|X) = \sum_{\pi \in \beta^{-1}(l)} \prod_{t=1}^{T} p(\pi_t|X). \tag{3}$$

In training, the parameters of the BLSTM $\boldsymbol{w}$ are optimized using back-propagation through time (BPTT) [17] by minimizing the following objective function:

$$E(\boldsymbol{w}) = -\log \sum_{\pi \in \beta^{-1}(l)} \prod_{t=1}^{T} p(\pi_t|X). \tag{4}$$

The gradient of the objective function can be efficiently calculated with the forward-backward algorithm.

During conversion, the output sequence $\hat{l}$ is estimated with respect to the given acoustic feature sequence X using best path decoding as follows:

$$\hat{l} = \beta(\arg\max_{\pi} p(\pi|X)). \tag{5}$$

V. Experimental evaluation

A. Experimental conditions

We divide our database of typical onomatopoeia transcriptions, constructed as described in Section III, into three groups of 9,120 samples, 500 samples, and 100 samples to form a training set, a validation set, and an evaluation set, respectively. The onomatopoeic transcriptions are converted from katakana character sequences into phoneme sequences. A phoneme set is composed of vowels and consonants, with the exception of geminate consonants, which are represented by "q", long vowels, which are represented by ":" (e.g. "a:"), and syllabic nasal sounds which are represented by "N."

We then conduct objective and subjective evaluations to investigate the effectiveness of the proposed onomatopoeia conversion method. In the objective evaluation, we measure conversion accuracy using two evaluation metrics, a word error rate (WER) and a phoneme error rate (PER), which are calculated as follows:

$$\mathrm{WER} = \frac{N_{\mathrm{correct}}}{N_{\mathrm{word}}}, \tag{6}$$

ISBN 978-90-827970-1-5

TABLE I
EXPERIMENTAL CONDITIONS

# layers	3
Window size	40 ms
Shift size	20 ms
BLSTM cell	512
Learning rate	0.001
Initial scale	0.001
Gradient clipping norm	5
Optimizer	Adam [18]
Time steps	350
Batch size	128
Epochs	20

TABLE II
RESULTS OF OBJECTIVE EVALUATION

	WER [%]	PER [%]
CTC	46.00	20.49
GMM-HMM	100.00	111.28

$$\mathrm{PER} = \frac{S + I + D}{N_{\mathrm{phoneme}}}, \tag{7}$$

where N_{correct} is the number of onomatopoeic representations correctly converted from the sound events, and N_{word} is the total number of the onomatopoeic representations in the evaluation set. S, I, and D correspond to the number of substitution errors, insertion errors, and deletion errors in edit distance, respectively. N_{phoneme} is the total number of phonemes included in the onomatopoeic representations in the evaluation set.

In the subjective evaluation, the onomatopoeic representations of the sound events are evaluated on the basis of whether or not they are accepted by the subjects. After each sound event sample is presented to the subject, the subject first transcribe it into an onomatopoeic representation to confirm the most suitable representation for him or her. After that, the onomatopoeic representation to be evaluated is presented to the subject, and he or she indicate whether or not it is an acceptable onomatopoeic representation for the corresponding sound event sample. Eight subjects, Japanese men and women in their twenties, participated in the evaluation.

The training conditions for the proposed sound event encoder model are shown in Table I. These settings are determined using a grid search so that the best possible performance is obtained using the objective evaluation metrics.

B. Experimental results of objective evaluation

As a reference for comparison, we also develop a sound event encoder based on GMM-HMM [19] using a Japanese phoneme recognition network. The results of the objective evaluation of the proposed and comparison methods are shown in Table II.

In comparison with the GMM-HMM-based sound event encoder, both WER and PER are significantly reduced when using the proposed method. Examples of some of the onomatopoeic transcriptions of several sound events using each method are shown in Table III. Unnatural Japanese onomatopoeic representations (e.g., phoneme sequences including redundant consecutive vowels) are observed in the results of the GMM-HMM-based sound event encoder. On the other hand, these errors are well suppressed by the proposed sound event encoder thanks to the use of CTC, which allows direct modeling sequence conversion processing.

TABLE III
EXAMPLES OF CONVERTED ONOMATOPOEIA

Ground truth	CTC	GMM-HMM
p i p o N	p i p o N	d a: hy a r a p u z u:
sh a r a r a r a	sh a r a r a	ch i h i h i u u u N
k a ch a:	k o t a k a N	o n i e e r a a a N
k o: N	k o: N	k i d u g u d u ky u ky a
p u k i N	k o N	n i g a g a p e i i
ch i N	ch i N	u ts u
gy u: N	gy u: N	a h e q
b a q	b a q	p u
t o N	t o N	n i n u a a q
j i:	j i j i j i j i j i j i	p a N
j i r i j i r i	j i r i j i r i	r e N
ch i q	ch i	a

TABLE IV
RESULTS OF SUBJECTIVE EVALUATION

Transcriptions	Acceptable [%]	Unacceptable [%]
One subject OD	74.0	26.0
CTC-OD	74.5	25.5
CTC-TOD	**81.8***	**18.2***
Self-labeled	91.3	8.7

C. Experimental results of subjective evaluation

The results of the subjective evaluation experiment are shown in Table IV. OD represents acceptance results when using the onomatopoeic representations from the original, one-person onomatopoeia database. The acceptance rate is 74.0%, i.e., only 74.0% of the onomatopoeic representations defined by a single individual are accepted by the other subjects. On the other hand, CTC-OD shows the results when using the onomatopoeic representations converted from sound event samples using the proposed encoder (developed using the original, single individual onomatopoeic representations). The acceptability rate is 74.5%, which is almost the same as the OD results. This confirms that the transcription accuracy of the proposed encoder is sufficiently high, about the same as human transcription. When we used the onomatopoeia database developed through several revisions by multiple subjects, the acceptability rate of the transcriptions of the proposed sound event encoder (CTC-TOD) improved to 81.8%, i.e., a 7.3% improvement, which is a statistically significant ($p < 0.05$). Therefore, the use of "typical" onomatopoeic representations is more effective when transcribing sound events.

To investigate the ambiguity of onomatopoeic representations among different individuals, examples of several onomatopoeic representations of the same sound events, as manually transcribed by five subjects, are shown in Table V. Al-

ISBN 978-90-827970-1-5 © EURASIP 2018

TABLE V
ONOMATOPOEIA TRANSCRIPTIONS OF FIVE SUBJECTS

Subject A	Subject B	Subject C	Subject D	Subject E
p i N p o: N	p i p o N	t e r e N	p i k o: N	p i p o N
sy a r a r a r a	ch i r i r i r i N	ch i r i N ch i r i N	r i N r i N	sy a r a r a r a
k a N p a t a	ch i N r i N	t a q t a r a N	k a N k a r a N	k a ch a:
t o t o t o t o	k o k o q k o k o q	t e q t e r e N	k a r a q k a r a N	k a r a k a r a
k a N	k o N	t a: N	ch i N	k o: N

though the onomatopoeic representations of the same sounds events differ, similar tendencies are observed among them, e.g., the number of morae is similar, and syllables at the ends of the transcriptions are also similar. It is expected that these characteristics, observed among multiple individuals, need to be modeled in order to define typical onomatopoeic representations. Moreover, to clarify an upper bound of conversion accuracy, we conduct another subjective evaluation by asking the subjects if their own onomatopoeic representations are acceptable. This evaluation is conducted two months after they have done the transcriptions. The average result is 91.3%, shown as "Self-labeled" in Table IV. We have also found that complex sound events, such as the sound of paper being crumpled, are relatively difficult to transcribe onomatopoeically.

Our results suggest that the proposed sound event encoder is capable of converting most sound events into acceptable onomatopoeic representations, although there remains room for further improvement.

VI. CONCLUSIONS

In this paper, we have proposed a connectionist temporal classification-based sound event encoder for transcribing sound events into onomatopoeic representations. The proposed method is an end-to-end approach, meaning it is possible to apply it to any waveform without preprocessing. In order to convert sound events into more acceptable onomatopoeic representations, we developed a database of typical onomatopoeic transcriptions by having multiple subjects double-check each other's transcriptions. Objective and subjective evaluation results demonstrated that the proposed method is capable of converting sound events into acceptable onomatopoeic representations and that having multiple subjects review the transcriptions included in the database yielded significant improvements.

In future work, we plan to expand our onomatopoeia database so that our proposed method will be able to transcribe a wider variety of sound events. Furthermore, we plan to implement a conversion system which can generate multilingual onomatopoeic representations.

ACKNOWLEDGMENT

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant Number 17H01763.

REFERENCES

[1] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. L. Roux, and K. Takeda, "Duration-controlled LSTM for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2059–2070, Nov 2017.

[2] M. Xu, L.-Y. Duan, J. Cai, L.-T. Chia, C. Xu, and Q. Tian, "Hmm-based audio keyword generation," vol. 3333, pp. 566–574, 11 2004.

[3] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Audio context recognition using audio event histograms," in *2010 18th European Signal Processing Conference*, Aug 2010, pp. 1272–1276.

[4] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Proceedings of the 2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, ser. AVSS '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 21–26.

[5] Y. Oishi and K. Tatsuta, "Physical characteristics of sound and onomatopoeic expressions in japanese : expressions for pure tones," *The Journal of the Acoustic Society of Japan*, vol. 72, no. 3, pp. 105–114, 2016.

[6] G. Lemaitre and D. Rocchesso, "On the effectiveness of vocal imitations and verbal descriptions of sounds," *The Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 862–873, 2014.

[7] S. A. Guynes, "Four-Color Sound: A Peircean Semiotics of Comic Book Onomatopoeia," *The Public Journal of Semiotics*, vol. 6, no. 1, pp. 58–72, 2013.

[8] S. Wake, "Sound retrieval with intuitive verbal descriptions," *IEICE TRANSACTIONS on Information and Systems*, vol. E84-D, no. 11, pp. 1568–1576, 2001.

[9] K. Ishihara, F. Kimura, and A. Maeda, "Music retrieval using onomatopoeic query," in *Proc. World Congress on Engineering and Computer Science (WCECS)*, 2013.

[10] F. Wang, H. Nagano, K. Kashino, and T. Igarashi, "Visualizing video sounds with sound word animation to enrich user experience," *IEEE Transactions on Multimedia*, vol. 19, no. 2, pp. 418–429, 2017.

[11] K. Ishihara, Y. Tsubota, and H. Okuno, *Automatic transformation of environmental sounds into sound-imitation words based on Japanese syllable structure*. International Speech Communication Association, 2003, pp. 3185–3188.

[12] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition." in *LREC*, 2000.

[13] M. Schuster and K. K. Paliwal, "Bidirectional Recurrent Neural Networks," *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, vol. 45, no. 11, pp. 2673–2681, 1997.

[14] S. Hochreiter and J. Urgen Schmidhuber, "LONG SHORT-TERM MEMORY," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, 2014.

[16] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." in *Interspeech*, 2014, pp. 338–342.

[17] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, pp. 1–15, 2014.

[19] A. Lee, T. Kawahara, and K. Shikano, "Julius - an open source real-time large vocabulary recognition engine." in *Proceedings of Eurospeech*, Aalborg, Denmark, 2001, pp. 1691–1694.

ISBN 978-90-827970-1-5 © EURASIP 2018