

# Contents

## Invited Talks

Speech Analysis in the Big Data Era . . . . .	3
<i>Björn W. Schuller</i>	

## Conference Papers

A Multi-criteria Text Selection Approach for Building a Speech Corpus . . . .	15
<i>Chiragkumar Patel and Sunil Kumar Kopparapu</i>	
Experiment with GMM-Based Artefact Localization in Czech Synthetic Speech. . . . .	23
<i>Jiří Přibíl, Anna Přibilová, and Jindřich Matoušek</i>	
Tuned and GPU-Accelerated Parallel Data Mining from Comparable Corpora . . . . .	32
<i>Krzysztof Wolk and Krzysztof Marasek</i>	
Investigating Genre and Method Variation in Translation Using Text Classification . . . . .	41
<i>Marcos Zampieri and Ekaterina Lapshinova-Koltunski</i>	
Extracting Characteristics of Fashion Models from Magazines for Item Recommendation . . . . .	51
<i>Taishi Murakami, Yoshiaki Kurosawa, Yuri Kurashita, Kazuya Mera, and Toshiaki Takezawa</i>	
Segment Representations in Named Entity Recognition . . . . .	61
<i>Michal Konkol and Miloslav Konopík</i>	
Analyzing Text Coherence via Multiple Annotation in the Prague Dependency Treebank . . . . .	71
<i>Kateřina Rysová and Magdaléna Rysová</i>	
Automatic Detection of Parkinson's Disease in Reverberant Environments . . .	80
<i>Juan Rafael Orozco-Arroyave, Tino Haderlein, and Elmar Nöth</i>	
Automatic Detection of Parkinson's Disease from Compressed Speech Recordings. . . . .	88
<i>Juan Rafael Orozco-Arroyave, Nicanor García, Jesús Francisco Vargas-Bonilla, and Elmar Nöth</i>	

Time Dependent ARMA for Automatic Recognition of Fear-Type Emotions in Speech. . . . .	96
<i>J.C. Vásquez-Correa, J.R. Orozco-Arroyave, J.D. Arias-Londoño, J.F. Vargas-Bonilla, L.D. Avendaño, and Elmar Nöth</i>	
Using Lexical Stress in Authorship Attribution of Historical Texts . . . . .	105
<i>Lubomir Ivanov and Smiljana Petrovic</i>	
Eye Gaze Analyses in L1 and L2 Conversations: Difference in Interaction Structures. . . . .	114
<i>Koki Ijuin, Yasuhiro Horiuchi, Ichiro Umata, and Seiichi Yamamoto</i>	
Word Categorization of Corporate Annual Reports for Bankruptcy Prediction by Machine Learning Methods . . . . .	122
<i>Petr Hájek and Vladimír Olej</i>	
Novel Multi-word Lists for Investors' Decision Making. . . . .	131
<i>Renáta Myšková and Petr Hájek</i>	
Topic Classifier for Customer Service Dialog Systems. . . . .	140
<i>Manex Serras, Naiara Perez, M. Inés Torres, Arantza Del Pozo, and Raquel Justo</i>	
Defining a Global Adaptive Duration Target Cost for Unit Selection Speech Synthesis . . . . .	149
<i>David Guennec, Jonathan Chevelu, and Damien Lolive</i>	
Adaptive Speech Synthesis of Albanian Dialects. . . . .	158
<i>Michael Pucher, Valon Xhafa, Agni Dika, and Markus Toman</i>	
Language-Independent Age Estimation from Speech Using Phonological and Phonemic Features . . . . .	165
<i>Tino Haderlein, Catherine Middag, Florian Hönig, Jean-Pierre Martens, Michael Döllinger, Anne Schützenberger, and Elmar Nöth</i>	
LecTrack: Incremental Dialog State Tracking with Long Short-Term Memory Networks . . . . .	174
<i>Lukáš Žilka and Filip Jurčiček</i>	
Automatic Construction of Domain Specific Sentiment Lexicons for Hungarian. . . . .	183
<i>Viktor Hangya</i>	

Investigation of Word Senses over Time Using Linguistic Corpora . . . . .	191
<i>Christian Pölit, Thomas Bartz, Katharina Morik, and Angelika Störrer</i>	
Dependency-Based Problem Phrase Extraction from User Reviews of Products. . . . .	199
<i>Elena Tutubalina</i>	
Semantic Splitting of German Medical Compounds . . . . .	207
<i>Claudia Bretschneider and Sonja Zillner</i>	
A Comparison of MT Methods for Closely Related Languages: A Case Study on Czech – Slovak and Croatian – Slovenian Language Pairs . . . . .	216
<i>Jernej Vičič and Vladislav Kuboň</i>	
Ideas for Clustering of Similar Models of a Speaker in an Online Speaker Diarization System . . . . .	225
<i>Marie Kunešová and Vlasta Radová</i>	
Simultaneously Trained NN-Based Acoustic Model and NN-Based Feature Extractor . . . . .	234
<i>Jan Zelinka, Jan Vaněk, and Luděk Müller</i>	
Named Entity Recognition for Mongolian Language . . . . .	243
<i>Zoljargal Munkhjargal, Gabor Bella, Altangerel Chagnaa, and Fausto Giunchiglia</i>	
Comparing Semantic Models for Evaluating Automatic Document Summarization . . . . .	252
<i>Michal Campr and Karel Ježek</i>	
Automatic Labeling of Semantic Roles with a Dependency Parser in Hungarian Economic Texts. . . . .	261
<i>Zoltán Subecz</i>	
Significance of Unvoiced Segments and Fundamental Frequency in Infant Cry Analysis . . . . .	273
<i>Anshu Chittora and Hemant A. Patil</i>	
Speech Corpus Preparation for Voice Banking of Laryngectomised Patients. . . . .	282
<i>Markéta Jůzová, Jan Romportl, and Daniel Tihelka</i>	
An Open Source Speech Synthesis Frontend for HTS . . . . .	291
<i>Markus Toman and Michael Pucher</i>	

Tibetan Linguistic Terminology on the Base of the Tibetan Traditional Grammar Treatises Corpus . . . . .	299
<i>Pavel Grokhovskiy, Maria Khokhlova, Maria Smirnova, and Victor Zakharov</i>	
Improving Multi-label Document Classification of Czech News Articles . . . .	307
<i>Jan Lehečka and Jan Švec</i>	
Score Normalization Methods for Relevant Documents Selection for Blind Relevance Feedback in Speech Information Retrieval . . . . .	316
<i>Lucie Skorkovská</i>	
Imbalanced Text Categorization Based on Positive and Negative Term Weighting Approach . . . . .	325
<i>Behzad Naderalvojoud, Ebru Akcapinar Sezer, and Alaettin Ucan</i>	
CloudASR: Platform and Service . . . . .	334
<i>Ondřej Klejch, Ondřej Plátek, Lukáš Žilka, and Filip Jurčíček</i>	
Experimental Tagging of the ORAL Series Corpora: Insights on Using a Stochastic Tagger . . . . .	342
<i>David Lukeš, Petra Klimešová, Zuzana Komrsková, and Marie Kopřivová</i>	
Development and Evaluation of the Emotional Slovenian Speech Database - EmoLUKS . . . . .	351
<i>Tadej Justin, Vitomir Štruc, Janez Žibert, and France Mihelič</i>	
A Semi-automatic Adjective Mapping Between plWordNet and Princeton WordNet . . . . .	360
<i>Ewa Rudnicka, Wojciech Witkowski, and Michał Kaliński</i>	
Toward Exploring the Role of Disfluencies from an Acoustic Point of View: A New Aspect of (Dis)continuous Speech Prosody Modelling. . . . .	369
<i>György Szaszák and András Beke</i>	
Heuristic Algorithm for Zero Subject Detection in Polish . . . . .	378
<i>Adam Kaczmarek and Michał Marcińczuk</i>	
Vocal Tract Length Normalization Features for Audio Search. . . . .	387
<i>Maulik C. Madhavi, Shubham Sharma, and Hemant A. Patil</i>	
RENA: A Named Entity Recognition System for Arabic . . . . .	396
<i>Ismail El bazi and Nabil Laachfoubi</i>	
Combining Evidences from Mel Cepstral and Cochlear Cepstral Features for Speaker Recognition Using Whispered Speech. . . . .	405
<i>Aditya Raikar, Ami Gandhi, and Hemant A. Patil</i>	

Random Indexing Explained with High Probability . . . . .	414
<i>Behrang QasemiZadeh and Siegfried Handschuh</i>	
Hungarian Grammar Writing with LTAG and XMG . . . . .	424
<i>Kata Balogh</i>	
Phonetic Segmentation Using KALDI and Reduced Pronunciation Detection in Causal Czech Speech . . . . .	433
<i>Zdenek Patc, Petr Mizera, and Petr Pollak</i>	
Automatic Robust Rule-Based Phonetization of Standard Arabic . . . . .	442
<i>Fadi Sindran, Firas Mualla, Katharina Bobzin, and Elmar Nöth</i>	
Knowledge-Based and Data-Driven Approaches for Georeferencing of Informal Documents . . . . .	452
<i>Daniel Ferrés and Horacio Rodríguez</i>	
Automated Mining of Relevant N-grams in Relation to Predominant Topics of Text Documents . . . . .	461
<i>Jan Žižka and František Dařena</i>	
Incremental Dependency Parsing and Disfluency Detection in Spoken Learner English . . . . .	470
<i>Russell Moore, Andrew Caines, Calbert Graham, and Paula Buttery</i>	
Open Source German Distant Speech Recognition: Corpus and Acoustic Model . . . . .	480
<i>Stephan Radeck-Arneth, Benjamin Milde, Arvid Lange, Evandro Gouvêa, Stefan Radomski, Max Mühlhäuser, and Chris Biemann</i>	
First Steps in Czech Entity Linking . . . . .	489
<i>Michal Konkol</i>	
Classification of Prosodic Phrases by Using HMMs . . . . .	497
<i>Zdeněk Hanzlíček</i>	
Adding Multilingual Terminological Resources to Parallel Corpora for Statistical Machine Translation Deteriorates System Performance: A Negative Result from Experiments in the Biomedical Domain . . . . .	506
<i>Johannes Hellrich and Udo Hahn</i>	
Derivancze — Derivational Analyzer of Czech . . . . .	515
<i>Karel Pala and Pavel Šmerk</i>	
Detection of Large Segmentation Errors with Score Predictive Model . . . . .	524
<i>Martin Matura and Jindřich Matoušek</i>	

Identification of Noun-Noun Compounds in the Context of Speech-to-Speech Translation . . . . .	533
<i>Maria Ivanova and Eric Wehrli</i>	
From Spoken Language to Ontology-Driven Dialogue Management . . . . .	542
<i>Dmitry Mouromtsev, Liubov Kovriguina, Yury Emelyanov, Dmitry Pavlov, and Alexander Shipilo</i>	
Entity-Oriented Sentiment Analysis of Tweets: Results and Problems . . . . .	551
<i>Natalia Loukachevitch and Yuliya Rubtsova</i>	
Improved Estimation of Articulatory Features Based on Acoustic Features with Temporal Context . . . . .	560
<i>Petr Mizera and Petr Pollak</i>	
Do Important Words in Bag-of-Words Model of Text Relatedness Help? . . . .	569
<i>Aminul Islam, Evangelos Milios, and Vlado Kešelj</i>	
Increased Recall in Annotation Variance Detection in Treebanks. . . . .	578
<i>Pablo Faria</i>	
Using Sociolinguistic Inspired Features for Gender Classification of Web Authors . . . . .	587
<i>Vasiliki Simaki, Christina Aravantinou, Iosif Mporas, and Vasileios Megalooikonomou</i>	
Modified Group Delay Based Features for Asthma and HIE Infant Cries Classification . . . . .	595
<i>Anshu Chittora and Hemant A. Patil</i>	
Self-Enrichment of Normalized LMF Dictionaries Through Syntactic- Behaviors-to-Meanings Links . . . . .	603
<i>Imen Elleuch, Bilel Gargouri, and Abdelmajid Ben Hamadou</i>	
<b>Author Index . . . . .</b>	<b>611</b>

Text, Speech, and Dialogue

18th International Conference, TSD 2015, Pilsen, Czech  
Republic, September 14-17, 2015, Proceedings

Král, P.; Matoušek, V. (Eds.)

2015, XVIII, 612 p. 122 illus. in color., Softcover

ISBN: 978-3-319-24032-9