

5th Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU-2016)

Procedia Computer Science Volume 81

Yogyakarta, Indonesia
9 - 12 May 2016

Editors:

**Sakriani Sakti
Mirna Adriani
Ayu Purwarianti**

**Laurent Besacier
Eric Castelli
Pascal Nocera**

ISBN: 978-1-5108-2397-6

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© by Elsevier B.V.
All rights reserved.

Printed by Curran Associates, Inc. (2016)

For permission requests, please contact Elsevier B.V.
at the address below.

Elsevier B.V.
Radarweg 29
Amsterdam 1043 NX
The Netherlands

Phone: +31 20 485 3911
Fax: +31 20 485 2457

<http://www.elsevierpublishingsolutions.com/contact.asp>

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Contents

SLTU2016 Preface	
S. Sakti, M. Adriani, A. Purwarianti, L. Besacier, E. Castelli, P. Nocera	1
Breaking the Unwritten Language Barrier: The BULB Project	
G. Adda, S. Stüker, M. Adda-Decker, O. Ambouroue, L. Besacier, D. Blachon, H. Bonneau-Maynard, P. Godard, F. Hamlaoui, D. Idiatov, G.-N. Kouarata, L. Lamel, E.-M. Makasso, A. Rialland, M. Van de Velde, F. Yvon, S. Zerbian	8
Study of Large Data Resources for Multilingual Training and System Porting	
F. Grézl, E. Egorova, M. Karafiát	15
Mismatched Crowdsourcing Based Language Perception for under-resourced Languages	
W. Chen, M. Hasegawa-Johnson, N.F. Chen	23
Performance Improvement of Probabilistic Transcriptions with Language-Specific Constraints	
X. Kong, P. Jyothi, M. Hasegawa-Johnson	30
Collaborative Speech Data Acquisition for under Resourced Languages through Crowdsourcing	
S. Arora, K.K. Arora, M.K. Roy, S.S. Agrawal, B.K. Murthy	37
Developing Speech Resources from Parliamentary Data for South African English	
F. de Wet, J. Badenhorst, T. Modipa	45
Eyra - Speech Data Acquisition System for Many Languages	
M. Petursson, S. Klüpfel, J. Gudnason	53
Parallel Speech Collection for Under-resourced Language Studies Using the LIG-AIKUMA Mobile Device App	
D. Blachon, E. Gauthier, L. Besacier, G.-N. Kouarata, M. Adda-Decker, A. Rialland	61
The Zero Resource Speech Challenge 2015: Proposed Approaches and Results	
M. Versteegh, X. Anguera, A. Jansen, E. Dupoux	67
Unsupervised Linear Discriminant Analysis for Supporting DPGMM Clustering in the Zero Resource Scenario	
M. Heck, S. Sakti, S. Nakamura	73
Variational Inference for Acoustic Unit Discovery	
L. Ondel, L. Burget, J. Černocký	80
Refining Sparse Coding Sub-word Unit Inventories with Lattice-constrained Viterbi Training	
W. Agenbag, T. Niesler	87
A Temporal Coherence Loss Function for Learning Unsupervised Acoustic Embeddings	
G. Synnaeve, E. Dupoux	95
Automatic Syllable Segmentation Using Broad Phonetic Class Information	
B. Ludusan, E. Dupoux	101
Lithuanian Broadcast Speech Transcription Using Semi-supervised Acoustic Model Training	
R. Lileikytė, A. Gorin, L. Lamel, J.-L. Gauvain, T. Fraga-Silva	107
Semi-Supervised Training of Language Model on Spanish Conversational Telephone Speech Data	
E. Egorova, J.L. Serrano	114
Automatic Speech Recognition of English-isiZulu Code-switched Speech from South African Soap Operas	
E. van der Westhuizen, T. Niesler	121
Code-switched English Pronunciation Modeling for Swahili Spoken Term Detection	
N. Kleynhans, W. Hartman, D. van Niekerk, C. van Heerden, R. Schwartz, S. Tsakalidis, M. Davel	128
Automatic Speech Recognition for African Languages with Vowel Length Contrast	
E. Gauthier, L. Besacier, S. Voisin	136
Bottle-Neck Feature Extraction Structures for Multilingual Training and Porting	
F. Grézl, M. Karafiát	144
Using Weighted Model Averaging in Distributed Multilingual DNNs to Improve Low Resource ASR	
R. Sahraeian, D. Van Compernelle	152
Investigating Bilingual Deep Neural Networks for Automatic Recognition of Code-switching Frisian Speech	
E. Yilmaz, H. van den Heuvel, D. van Leeuwen	159
Towards Robust Indonesian Speech Recognition with Spontaneous-Speech Adapted Acoustic Models	
D. Hoesen, C.H. Satriawan, D.P. Lestari, M.L. Khodra	167
The Effect of Tone Modeling in Vietnamese LVCSR System	
Q. Bao Nguyen, T. Thang Vu, C.M. Luong	174

Spoken Language Identification with Phonotactics Methods on Minangkabau, Sundanese, and Javanese Languages N.E. Safitri, A. Zahra, M. Adriani	182
A Tool to Solve Sentence Segmentation Problem on Preparing Speech Database for Indonesian Text-to-Speech System M.T. Uliniansyah, G.E. Nurfadhilah, L.R. Aini, J. Junde, F. Ayuningtyas, A. Santosa	188
Building Statistical Parametric Multi-speaker Synthesis for Bangladeshi Bangla A. Gutkin, L. Ha, M. Jansche, O. Kjartansson, K. Pipatsrisawat, R. Sproat	194
Automatic Technologies for Processing Spoken Sign Languages A. Karpov, I. Kipyatkova, M. Zelezny	201
Dictionary-based Word Segmentation for Javanese D. Tanaya, M. Adriani	208
Exploiting Syntactic Similarities for Preposition Error Corrections on Indonesian Sentences Written by Second Language Learner B. Irmawati, H. Shindo, Y. Matsumoto	214
Indonesian Named-entity Recognition for 15 Classes Using Ensemble Supervised Learning A.S. Wibawa, A. Purwarianti	221
Topic Summarization of Microblog Document in Bahasa Indonesia Using the Phrase Reinforcement Algorithm M.A. Jiwanggi, M. Adriani	229
The Development of an Audible Pattani Malay-Thai Electronic Phrasebook for Military Purposes P. Boonkwan, T. Supnithi, W. Tosuwan, C. Wutiwiwatchai	237
Using Dictionary and Lemmatizer to Improve Low Resource English-Malay Statistical Machine Translation System Y.-L. Yeong, T.-P. Tan, S.K. Mohammad	243
A Study of Statistical Machine Translation Methods for under Resourced Languages W. Pa Pa, Y.K. Thu, A. Finch, E. Sumita	250