Editorial

Introduction to the Special Issue on Processing Under-Resourced Languages

The creation of language and acoustic resources, for any given spoken language, is typically a costly task. For example, a large amount of time and money is required to properly create annotated speech corpora for automatic speech recognition (ASR), domain-specific text corpora for language modeling (LM), etc. The development of speech technologies (ASR, Text-to-Speech) for the already high-resourced languages (such as English, French or Mandarin, for example) is less constrained by this issue and, consequently, high-performance commercial systems are already on the market. On the other hand, for under-resourced languages, the above issue is typically the main obstacle.

Given this, the scientific community's concern with porting, adapting, or creating language and acoustic resources or even models for low-resourced languages has been growing recently and several algorithms and methods of adaptation have been proposed and experimented with. In the mean time, workshops and special sessions took place recently on this domain. For instance, the guest editors of this journal issue have organized a special session on under-resourced languages at Interspeech 2011. A bi-annual workshop called SLTU (Spoken Language Technologies for Under-resourced languages) will also have its 4th edition in 2014.

Apart from the survey paper by the guest editors, 23 submissions were received for this special issue in August 2012, out of which 11 were accepted as the result of 2 rounds of the reviewing process (48% acceptance rate): 6 papers deal with automatic speech recognition (ASR) in one or several under-resourced languages, 2 papers are dedicated to text-to-speech synthesis (TTS) and 3 papers deal with data collection or knowledge discovery for under-resourced and under-described languages. Moreover, several under-resourced languages of Asia (Indian languages), Africa (Yoruba, Amharic, Ibibio, languages from South-Africa) and Europe (Romanian) are dealt with in the different papers of this issue.

The review article « *Automatic Speech Recognition for Under-Resourced Languages: A Survey* » aims at providing a broad overview of the state of the art in the field of automatic speech recognition (ASR) for low-resourced languages. We believe that this paper will be a good starting point for anyone interested to initiate research in (or operational development of) ASR for one or several under-resourced languages. It should be clear, however, that many of the issues and approaches presented here, apply to speech technology in general (text-to-speech synthesis for instance).

Next come six articles dealing with general and multilingual topics in the field of automatic processing under-resourced languages:
In «*Web-based tools and methods for rapid pronunciation dictionary creation*» Tim Schlippe, Sebastian Ochs, and Tanja Schultz propose to use the World Wide Web as a seed for the rapid generation of pronunciation dictionaries in new languages. Language and vocabulary coverage of the Wiktionary resource is analyzed for multiple languages and statistical grapheme-to-phoneme models are built and evaluated in ASR for six different languages.
In «*A smartphone-based ASR data collection tool for under-resourced languages*», Nic J. de Vries, Marelie H. Davel, Jaco Badenhorst, Willem D. Basson, Febe de Wet, Etienne Barnard, and Alta de Waal report on a smartphone application that significantly simplifies the task of speech data collection. Although the tool has been found useful in practical data collection, it is interesting to note that the quality-control functionality of the tool does not seem to have a beneficial impact on the process.
In «*Eigentrigraphemes for under-resourced languages*» Tom Ko, and Brian Mak port the eigentriphone modeling method from a phone-based system to a grapheme-based system; the new method is called eigentrigrapheme modeling. Experiments on four official South African under-resourced languages (Afrikaans, South African English, Sesotho, siSwati) show the efficiency of the eigentrigrapheme modeling method.

In the article «*Using out-of-language data to improve an under-resourced speech recognizer*» by David Imseng, Petr Motlicek, Hervé Bourlard, and Philip N. Garner, it is reported on boosting the performance of an Afrikaans ASR system by using some Dutch data. The authors exploit multilingual resources through posterior features, estimated by multilayer perceptrons (MLP), and subspace Gaussian mixture models (SGMMs). 3 different acoustic modeling techniques (Tandem, Kullback–Leibler divergence based HMMs, and SGMMs) are studied, and the authors report on 12% relative improvement of the proposed multilingual system compared to a conventional monolingual HMM/GMM system trained on Afrikaans data only.

In «*Discovering the phoneme inventory of an unwritten language: a machine-assisted approach*» Timothy Kempton, Roger K. Moore suggest the use of advanced speech technologies to help field linguists in their work. More preceisely, they propose a machine-assisted approach for phonemic analysis of under-resourced and under-documented languages. Several procedures are investigated (phonetic similarity, complementary distribution, and minimal pairs) and compared.

In «*Acoustic modelling for speech recognition in Indian languages in an agricultural commodities task domain*» Aanchan Mohan, Richard Rose, Sina Hamidi Ghalehjegh, S. Umesh considers a small vocabulary speech recognition task in multiple Indian languages and proposes to use data from two linguistically similar languages – Hindi and Marathi — and to normalize the multilingual corpora shared with a variant of speaker adaptive training (SAT). The resulting multi-lingual system provides the best speech recognition performance for both languages.

The following three articles deal with new approaches for automatic speech recognition of different under-resourced languages:

In «*Using different acoustic, lexical and language modeling units for ASR of an under-resourced language – Amharic*» Martha Yifiru Tachbelie, Solomon Teferra Abate, Laurent Besacier propose hybrid phone/syllable based acoustic models for ASR of the Amharic language spoken in Ethiopia. Both word-based and morpheme-based language models are considered and the hybrid acoustic models bring a significant word error rate (WER) reduction compared to triphone-based systems for the best language model setting (morpheme-based).

In «*SMT-based ASR domain adaptation methods for under-resourced languages: Application to Romanian*» Horia Cucu, Andi Buzo, Laurent Besacier, Corneliu Burileanu investigate the possibility of using statistical machine translation to create domain-specific language resources. More precisely, the authors propose a methodology that aims to create a domain-specific automatic speech recognition (ASR) system for a low-resourced language when in-domain text corpora are available only in a high-resourced language. As a by-product of this core domain-adaptation methodology, this paper also presents the first large vocabulary continuous speech recognition system for Romanian.

In «*Large vocabulary Russian speech recognition using syntactico-statistical language modeling*» by Alexey Karpov, Konstantin Markov, Irina Kipyatkova, Daria Vazhenina, and Andrey Ronzhin, a novel approach for language model creation is proposed, which combines both syntactical and statistical analysis of available training text data. This approach shows advantages for LVCSR of Russian and other synthetic languages with a high freedom of language grammar (e.g., Slavic languages). Also the authors propose and study a combined knowledge-based statistical phoneme set selection method for obtaining an optimal set for ASR.

Finaly two articles deal with new approaches for speech synthesis of different under-resourced languages:

In «*Predicting utterance pitch targets in Yoruba for tone realisation in speech synthesis*» Daniel R. Van Niekerk, Etienne Barnard, features influencing syllable pitch targets in continuous utterances in Yorùbá are investigated in a small speech corpus of 4 speakers. It is found that the previous syllable pitch level is strongly correlated with pitch changes between syllables and a number of approaches and features are evaluated in this context. The resulting models can be used to predict utterance pitch targets for speech synthesisers (whether it be concatenative or statistical parametric systems), and may also prove useful in speech-recognition systems.

In «*Statistical parametric speech synthesis for Ibibio*» Moses Ekpenyong, Eno-Abasi Urua, Oliver Watts, Simon King, Junichi Yamagishi present the first statistical parametric speech synthesiser for Ibibio (spoken in Nigeria). The authors compare various representations of linguistic context and listening tests show that the use of tone marking contributes significantly to the quality of synthetic speech.

The guest editors are grateful to the Editors-in-Chief, Marc Swerts and then Bernd Möbius, as well as to 56 outstanding reviewers, who provided detailed and insightful reviews of the submissions: Martine Adda-Decker, Ebru Arisoy, Gerard Bailly, Frederic Bechet, Herve Blanchon, Herve Bourlard, Susanne Burger, Genevieve Caelen-Haumont, Khalid Choukri, Marelie Davel, Febe de Wet, Anoop Deoras, Christoph Draxler, Thierry Dutoit, Matthias Eck, Hynek Hermansky, Ruediger Hoffmann, Simon King, Daniil Kocharov, Aditi Lahiri, Javier Latorre, Viet Bac Le, Lin-Shan Lee, Haizhou Li, Georges Linares, Zhen-Hua Ling, Bin Ma, Konstantin Markov, Jean-Pierre Martens, Ludek Muller, Hema A Murthy, Seiichi Nakagawa, Udhyakumar Nallasamy, Thomas Richard Niesler, Pascal Nocera, Odetunji Odejobi, Ilya Oparin, Sathish Chandra Pammi, Thomas Pellegrini, Nitendra Rajput, Henning Reetz, Andrey Ronzhin, Leon Rothkrantz, Yoshinori Sagisaka, Germán Sanchis-Trilles, Murat Saraclar, Tim Schlippe, Sebastian Stüker, João Paulo Ramos Teixeira, Louis ten Bosch, Isabel Trancoso, Jurgen Trouvain, Liliya Tsirulnik, Dirk Van Compernolle, Daniel Rudolph Van Niekerk, and Christian Wellekens.

Editors
Laurent Besacier
/Laboratory of Informatics of Grenoble, Grenoble, France/
Etienne Barnard
/North-West University, Vanderbijlpark, South Africa/
Alexey Karpov
/St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia/
Tanja Schultz
/University of Karlsruhe, Karlsruhe, Germany/