Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 56 (2014) 85–100

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Automatic speech recognition for under-resourced languages: A survey

Laurent Besacier [a], Etienne Barnard [b], Alexey Karpov [c], Tanja Schultz [d]

[a] Laboratory of Informatics of Grenoble, Grenoble, France
[b] North-West University, Vanderbijlpark, South Africa
[c] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia
[d] Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Available online 7 August 2013

Abstract

Speech processing for under-resourced languages is an active field of research, which has experienced significant progress during the past decade. We propose, in this paper, a survey that focuses on automatic speech recognition (ASR) for these languages. The definition of under-resourced languages and the challenges associated to them are first defined. The main part of the paper is a literature review of the recent (last 8 years) contributions made in ASR for under-resourced languages. Examples of past projects and future trends when dealing with under-resourced languages are also presented. 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).

Keywords: Under-resourced languages; Automatic speech recognition (ASR); Language portability; Speech and language resources acquisition; Statistical language modeling; Crosslingual acoustic modeling and adaptation; Automatic pronunciation generation; Lexical modeling

1. Introduction

Nowadays, computers are heavily used to communicate via text and speech. Text processing tools, electronic dictionaries, and advanced speech processing systems like text-to-speech (speech generation) and speech-to-text (speech recognition) systems are readily available for several languages. There are however more than 6900 languages in the world and only a small fraction offers the resources required for implementation of Human Language Technologies (HLT). Thus, HLT are mostly concerned with languages for which large resources are available or which have suddenly become of interest because of the economic or political scene. Unfortunately, most languages from developing countries or minorities received only little attention so far. One way of improving this "language divide" is to do more research on the portability of speech and language technologies for multilingual applications, especially *for under-resourced languages*.

This paper is a review on automatic speech recognition (ASR) for under-resourced (UR) languages, which have shown a growing interest in the recent years. While the task of ASR is rather specific, some issues addressed in this paper apply to other HLT tasks as well. This paper is organized as follows: After an *Introduction* that focuses on the language diversity and on our motivation to address the topic, Section 2 gives a brief definition of what we call "under-resourced languages", as well as the challenges associated to them. Section 3 is a literature review of the recent contributions made in ASR for under-resourced languages. Examples of past projects on this topic are given in Section 4, while Section 5 presents the future trends when dealing with under-resourced languages. Finally, Section 6 concludes this work.

1.1. Languages of the world

Counting the number of languages in the world is not a straightforward task. First, one has to define what makes a language, for example to decide if dialects are considered to be a language, if so, which ones should be added, or if not, to draw the line between a language and a dialect. An

0167-6393/$ - see front matter © 2013 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.specom.2013.07.008

estimate for the total number of living languages in the world can be found on the Ethnologue[1] web site. They define a living language as "one that has at least one speaker for whom it is their first language". So, extinct languages and languages that are spoken as a second language are excluded from these counts. Based on this definition, Ethnologue lists 6909 known living languages. This list includes 473 languages that are classified as nearly extinct, i.e. when "only a few elderly speakers are still living". It is important to note that Ethnologue's list includes both verbal and visual-kinetic spoken languages. The latter ones are known as sign languages, which are used for everyday communication by the deaf; these spoken languages combine hand gestures with lips articulation and facial mimics. Almost all countries over the world define their own national sign languages.

Counting how many languages have a written form is also subject to controversy. The foundation for endangered languages web site[2] mentions 2000 written languages by counting published bibles (entirely or portions) but this also includes non-living languages. Omniglot,[3] an online encyclopedia of writing systems and languages, lists less than 1000 written languages and gives details on more than 180 different writing systems.

While counting languages is a tricky task, the number of "well-resourced languages" can be easily given by listing how many languages are identified for core technologies and resources, such as: Google Translate (63 languages involved[4] in 2012), Google search (more than one hundred languages in 2012), Siri ASR application (8 languages in 2012), Wiktionary[5] (~80 languages in 2012), Google Voice Search (29 languages and accents in 2012).

1.2. Language extinction

In today's globalized world, languages are disappearing at an alarming rate. Crystal (2000) estimated that over the next century about half of all existing languages will be extinct. On average, one could say that every two weeks one language dies. A survey by the Summer Institute of Linguistics (SIL) from February 1999 revealed that about 51 languages are left with only one speaker, 500 languages have 500 speakers left, and 3000 languages have less than 10.000 speakers left. The graph below summarizes the estimates of speakers over languages from the SIL survey. It shows that 96% of the world's languages are spoken by only 4% of its people.

History has shown that not even a language with 100.000 remaining speakers is safe from extinction (Crystal, 2000). The survival of a language depends on the pressure imposed on that language and on its speakers. Pressure may arise from disasters (earthquakes on Papua New Guinea killed several languages), genocide (about 90% of America's natives died within 200 years of European conquering) or simply from the dominance of another language. The latter may result in cultural assimilation (social, political or economic benefits to speak the dominant language) that usually leads to the loss of the suppressed language within few generations (e.g. second generation immigrants).

How could language extinction be slowed down and what are the associated costs. First of all, a language can only be saved if the community itself wants it and the surrounding culture respects this wish. Typically, the community is then supported to fund courses, materials, and teachers. In addition, linguists go into the field, collect and publish language related information such as grammars, dictionaries, speech recordings, and make them available to the public at large. The associated costs depend on the particular conditions, for example if the language has a writing system, etc. Crystal estimates about USD 80.000 per year per language. Considering 3000 endangered languages this would add up to more than USD 700 Million. Organizations like the Foundation of Endangered Languages (FEL) and large-scale UNESCO projects have been established to raise both, attention and funds, to tackle this major challenge (see Fig. 1).

1.3. Good reasons to address less prevalent languages

Some languages might be more attractive than others for Human Language Technologies (HLT). However, for the reasons described above, there are good reasons for developing speech recognition (and other technologies like machine translation) systems for literally all languages in the world. First of all, spoken language is the primary means of human communication. Both, individual and community memories, ideas, major events, practices, and lessons learned are all preserved and transmitted through language. Furthermore, language is not only a communication tool but fundamental to cultural identity and empowerment. So, language diversity in the world is the basis of our rich cultural heritage and diversity. If the world loses a language, the memories and experiences of this culture go with it. Crystal claims that language diversity should be treated like bio-diversity as history has shown that the more diverse eco-systems are strongest.

Human Language Technologies have a lot to offer to revitalize and (at least) document languages and thus prevent or slow down language extinction. The existence of technology may raise interest and make the language attractive again to their native speakers. Moreover, in the perspective of saving some endangered languages (some mostly spoken and not written), the possibility to rapidly develop ASR systems to transcribe them is an important step for their preservation and would facilitate access to audio contents in these languages. A second reason why

[1] http://www.ethnologue.com/
[2] http://www.ogmios.org/home.htm.
[3] http://www.omniglot.com.
[4] http://www.techcentral.co.za/googles-babel-fish-heralds-future-of-translation/28396/.
[5] http://www.wiktionary.org/.

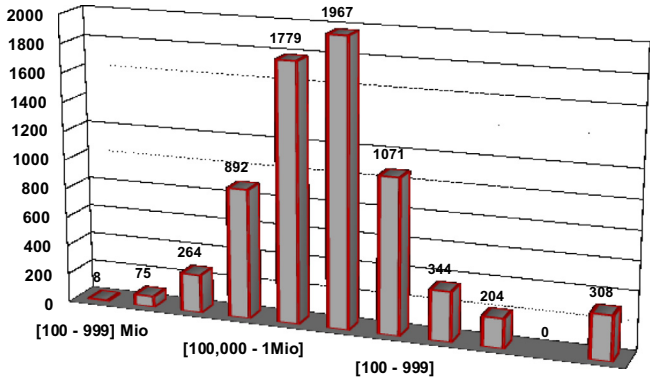

Fig. 1. Graph of SIL survey about languages extinction.

HLT should be available for all languages is that the political impact of a language can be very volatile. In today's world, language is one of the few remaining barriers that hinder human-to-human interaction. Events such as armed conflicts or natural disasters might make it important to be able to communicate with speakers of a less-prevalent language, e.g. for humanitarian workers in a disaster area (see, for instance, the earthquake in Haiti that highlighted the need for technologies to handle Haitian Creole language[6]). Often, the people that one need to communicate with in such a scenario only speaks their own language that is unknown to the outsider, e.g. a foreign doctor trying to help. For these cases, human translators are often not available in necessary numbers and in a timely manner. Here, readily available technology such as speech translation systems can be highly beneficial. Such technology might be far from being perfect, but when being faced with the alternative of having no translation system at all for an unknown language in an emergency situation, the imperfect system will be of great use. Therefore, HLT should be developed especially for under-resourced languages. Last but not least, some under-resourced languages may blossom in the future to become of very strong social, political, or economic power (see for instance languages from rapidly developing countries, such as: Bengali, Malay, Vietnamese, Urdu or vehicular languages from Africa – Swahili, Wolof – some of them already being in the top-20 of the most spoken languages in the world).

2. Under-resourced (UR) languages

2.1. Definition

The term "under-resourced languages" introduced by Krauwer (2003) and Berment (2004) refers to a language with some of (if not all) the following aspects: lack of a unique writing system or stable orthography, limited presence on the web, lack of linguistic expertise, lack of electronic resources for speech and language processing, such as monolingual corpora, bilingual electronic dictionaries, transcribed speech data, pronunciation dictionaries, vocabulary lists, etc. The synonyms for the same concept are: low-density languages, resource-poor languages, low-data languages, less-resourced languages. It is important to note that it is not the same as a minority language which is a language spoken by a minority of the population of a territory. Some under-resourced languages are actually official languages of their country and spoken by a very large population. On the other hand, some minority languages can be considered as rather well-resourced (see for instance Catalan language available for Google Search and Google Translate). Consequently, under-resourced languages are not necessarily endangered (while the opposite is usually true).

2.2. Measure the status of a language

In order to objectively define the status of a language, the concept of BLARK (Basic LAnguage Resource Kit[7]) was defined in a joint initiative between European Network of Excellence in Language and Speech (ELSNET) and European Language Resources Association (ELRA) (Krauwer, 2003). From this project, a minimal set of language resources, to be made available for as many languages as possible, was defined. A similar matrix was presented in Berment (2004): a list of services is evaluated for a given language by an expert and a mean score is calculated (marks for each service are weighted by the criticity or importance of the service). Berment (2004) gives an example of this metric applied to Khmer, a language mainly spoken in Cambodia (6.2/20). The same metric evaluated for Vietnamese the same year gives 10/20. An under-resourced language is defined as a language which has a score below 10/20. More recently, METANET (a Network of Excellence consisting of 60 research centers from 34 countries) produced a series of white papers[8] entitled "Languages in the European Information Society" which report on the state of each European language with respect to Language Technology and explains the most urgent risks and chances. The key results show that some European languages are still considered as under-resourced[9] (for speech processing, the following languages are mentioned: Croatian, Icelandic, Latvian, Lithuanian, Maltese and Romanian).

2.3. Challenges

Porting HLT system (e.g. a speech recognition system) to an under-resourced language requires techniques that go far beyond the basic re-training of the models. Indeed,

[6] http://research.microsoft.com/apps/video/dl.aspx?id=136704 (Jeff Allen seminar in 2010).

[7] http://www.blark.org/.

[8] http://www.meta-net.eu/whitepapers/overview.

[9] http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison.

processing a new language often leads to new challenges (special phonological systems, word segmentation problems, fuzzy grammatical structure, unwritten language, etc.). The lack of resources requires, on its side, innovative data collection methodologies (via crowdsourcing for instance, see (Gelas et al., 2011)) or models for which information is shared between languages (e.g. multilingual acoustic models (Schultz, 2006; Schultz and Waibel, 2001; Le and Besacier, 2009)). In addition, some social and cultural aspects related to the context of the targeted language bring additional problems: languages with many dialects in different regions, code-switching or code-mixing phenomena (switching from one language to another within the discourse), massive presence of non-native speakers (in vehicular languages such as Swahili).

Finally, one has to bridge the gap between language experts (the speakers themselves) and technology experts (system developers). Indeed, it is often almost impossible to find native speakers with the necessary technical skills to develop ASR systems in their native language. Moreover, under-resourced languages are often poorly addressed in the linguistic literature and very few studies describe them. To bootstrap systems for such languages, one has to borrow resources and knowledge from similar languages, which requires the help of dialectologists (find proximity indices between languages), phoneticians (map the phonetic inventories between the targeted under-resourced language and some more resourced ones, etc.). Moreover, for some languages, it is sometime interesting to challenge the paradigms and common practices: is the word the best unit for language modeling? Is the phoneme the best unit for acoustic modeling? In addition, for some (rare, endangered) languages, it is often necessary to work with ethno-linguists in order to access to native speakers and in order to collect data in accordance with the basic technical and ethical rules. All of these aspects make research on technologies for under-resourced languages, a multi-disciplinary challenge.

2.4. Short history on under-resourced language research

In the nineties, ASR systems developed originally for one language had been successfully ported to other languages, including systems developed by IBM (Cohen et al., 1997), Dragon (Barnett et al., 1996), BBN (Billa et al., 1997), Cambridge (Young et al., 1997), Philips (Dugast et al., 1995), MIT (Glass et al., 1995), and LIMSI (Lamel et al., 1995). The transformation of English systems to such diverse languages like German, Japanese, French, and Mandarin Chinese illustrated that speech technology generalizes across languages and that similar modeling assumptions hold for various languages. In the late nineties, researchers started to systematically investigate the fitness of language independent acoustic models to bootstrap unseen languages. Studies looked at the impact of language families (Constantinescu and Chollet, 1997), the impact of the amount of languages used to create acoustic models (Gokcen and Gokcen, 1997; Schultz and Waibel, 1998; Schultz et al., 2007), the impact of the amount of training data (Wheatley et al., 1994; Köhler, 1998) and the question on how to share acoustic models across languages (Schultz and Waibel, 1998; Köhler, 1998). One of the early findings was that multilingual acoustic models outperform monolingual ones for the purpose of rapid language adaptation (Schultz and Waibel, 2001).

In the last 7 years, the scientific community's concern with porting, adapting, or creating written and spoken resources or even models for low-resourced languages has been growing. For instance, several adaptation methods have been proposed and experimented with lately, while workshops and special sessions have been organized on this issue. For instance, the workshop on Spoken Language Technologies for Under-resourced Languages (SLTU) took place in 2008 (Hanoi, Vietnam), 2010 (Penang, Malaysia) and 2012 (Cape Town, South Africa). In addition, a special session on Speech Technology for Under-Resourced Languages was held during the Interspeech 2011 conference.[10] While these events concerned languages from various places (at SLTU 2012[11], 17 languages from four different continents were addressed), some recent LREC or COLING workshops are now specific to geographic areas (see for instance Workshop on Indian Language Data: Resources and Evaluation; Workshop on Language Resources & Technologies for Turkic Languages; Workshop on Parsing in Indian Languages; Workshop on South and Southeast Asian Natural Language Processing, etc.).

2.5. Language resources

As described in detail below, the building process of ASR systems requires transcribed speech recordings from many speakers, pronunciation dictionaries which cover the full vocabulary of at least the training corpus, and massive amounts of text data to reliably train statistical language models.

While the amount of languages for which large-scale speech and text data resources that have been systematically collected and distributed has been growing during recent years, it still to-date does not cover more than about 100 languages (compared to about 50 languages 5 years ago). The Linguistic Data Consortium (LDC) has managed the design and collection of numerous large databases for the latest languages-of-interest, and provides corpora for ASR in many domains and conditions. The European Language Resources Association (ELRA) also provides databases in multiple languages with an emphasis on European languages. Other providers like AppenButlerHill list about 80 languages in their catalog[12] (accessed in July

[10] SLTU and Interspeech-2011 special session were organized or co-organized by the authors of this paper.
[11] http://www.mica.edu.vn/sltu2012/.
[12] http://catalog.appenbutlerhill.com/.

2013) and SpeechOcean provides databases in around 35 languages for ASR[13] (in July 2013). Nevertheless, the collection of databases in many regions is met with political and cultural barriers and the cost of licensing databases in certain languages might be prohibitive, especially for commercial companies. In addition, when it comes to pronunciation dictionaries and large text collections, the amount of languages is markedly smaller.

While we expect the amount of language to grow further, surprisingly few data collections emphasize uniform collection scenarios across languages. Such collections are expected to provide data of many languages with same recording quality (sampling rate, microphone type, noise conditions), speaking styles (read, conversational), transcription and dictionary formats, and of same domains. Such databases are required to train multilingual models which – a shared view within the community – are very useful for rapid portability to new languages and domains. One of the few exceptions is GlobalPhone, a standardized multilingual text and speech database (Schultz, 2002). This data collection provides transcribed speech data for the development and evaluation of multilingual spoken language processing systems in the most widespread languages of the world. GlobalPhone is designed to be uniform across languages with respect to the amount of text and speech per language (100 speakers per language), the audio quality (microphone, noise, channel), the collection scenario (task, setup, speaking style etc.), as well as the transcription and phone set conventions. As a consequence, GlobalPhone supplies an excellent basis for research in the areas of (1) multilingual speech recognition, (2) rapid deployment of speech processing systems to yet unsupported languages, (3) language identification tasks, (4) speaker recognition in multiple languages, (5) multilingual speech synthesis, as well as (6) monolingual speech recognition in a large variety of languages. To date, GlobalPhone covers 21 languages, including Arabic (MSA), Bulgarian, Chinese–Mandarin, Chinese–Shanghai, Croatian, Czech, French, German, Hausa, Japanese, Korean, Polish, Portuguese (Brazilian), Russian, Spanish (Latin American), Swedish, Tamil, Thai, Turkish, Ukrainian, and Vietnamese. In total the corpus contains over 400 h of speech spoken by more than 2000 native adult speakers (Schultz et al., 2013), together with pronunciation dictionaries, and freely accessible language models[14] to benchmark ASR systems in many languages.

Recent years have also seen the release of various corpora for the Southern African languages, including the relatively small AST (Roux et al., 2000) and Lwazi (Barnard et al., 2009) corpora of telephone speech, and the substantially larger NCHLT corpus (containing broadband speech) (De Vries et al., 2013). These corpora are all focused on the eleven official languages of South Africa, but the same or closely related languages are spoken in several Southern African countries.

3. Automatic speech recognition for under-resourced languages (U-ASR)

3.1. Components of ASR systems

Automatic speech recognition (ASR) converts a speech signal into a textual representation, i.e. sequence of said words by means of an algorithm implemented as a software or hardware module. Several types of natural speech and corresponding ASR systems are identified: spelled speech (with pauses between letters or phonemes), isolated speech (with pauses between words), continuous speech (when a speaker does not make any pauses between words), spontaneous speech (e.g. in a human-to-human dialog), and highly conversational speech (e.g. meetings and discussions of several people). ASR systems can be classified by the recognition vocabulary/lexicon size (Whittaker and Woodland, 2001): small (up to thousand words), medium (up to 10 K words), large (up to 100 K words), very/extra large (>100 000 words that is adequate for ASR for synthetic inflective and agglutinative languages and large domains; for instance 800 K words for Arabic), unlimited vocabulary (attempts to model all potential words of a language). Modern automatic speech recognizers are built using various techniques, such as Hidden Markov Models (HMM) (Young et al., 2008), Dynamic Time Warping (DTW) or Dynamic Programming (Jing et al., 2010), Dynamic Bayesian Networks (DBN) (Stephenson et al., 2002), Support Vector Machines (SVM) (Solera-Urena et al., 2007) or some hybrid models (Trentin and Gori, 2001; Ganapathiraju et al., 2000). Artificial Neural Networks (ANN) including single hidden layer NN and multiple hidden layers NN (Deep Neural Networks DNN or Deep Belief Networks DBN) are also used for ASR subtasks such as acoustic modeling (Mohamed et al., 2012; Seide et al., 2011) and language modeling (Arisoy et al., 2012; Mikolov et al., 2010).

General architecture of a standard ASR system that uses the stochastic HMM-based approach is presented in Fig. 2; it integrates three main components (Young et al., 2008): acoustical (acoustic–phonetic) modeling, lexical modeling (pronunciation lexicon/vocabulary) and language modeling. Any state-of-the-art ASR system works in two modes: model training and speech decoding. Purpose of the system training process is to create and improve models for speech acoustics (recordings of a lot of speakers are required for speaker-independent ASR), language (a corpus of training text data or sentence grammar is needed) and recognition lexicon (a list of the recognizable tokens with single or multiple phonetic transcriptions). Acoustical modeling allows representing the audio signals discriminating classes of basic speech units (context-independent such as monophones, syllables or context-dependent such as allophones, triphones, pentaphones) and taking into account speech

[13] http://www.speechocean.com/en-Product-Catalogue/.
[14] http://csl.ira.uka.de/GlobalPhone.

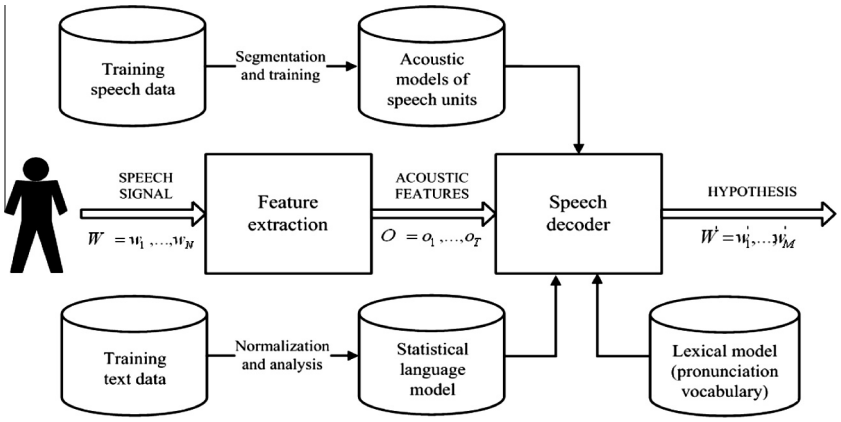

Fig. 2. Architecture of a state-of-the-art automatic speech recognition system and its components.

variability with respect to the speakers, channel, and environment. Vectors of speech signal features (e.g., mel-frequency cepstral coefficients (MFCC), linear prediction coefficients (LPC), perceptual linear prediction coefficients (PLP), bottleneck features (ML), etc.) are extracted from the acoustical signal for dimensionality reduction and probabilistic modeling. Lexical modeling aims at generating the recognition vocabulary and assigning each orthographic token (words or sub-words) of the lexicon with the corresponding spoken representation (phonetic transcription). Language modeling is needed to impose the constraints on recognition hypotheses generated during ASR and to model the structure, syntax and semantics of the target language. Statistical language models are based on the empirical fact that a good estimation of the probability of a lexical unit can be obtained by observing it on large text data.

Any ASR system integrates a speech decoder, which performs speech input processing and converting audio speech signals into a sequence of orthographic words. HMM-based speech decoders are usually based on the token passing method based on the Viterbi algorithm (Young et al., 2008). State-of-the-art speech decoders are able to generate word/phoneme *N*-best lists or lattices as a compact representation of the recognition hypotheses and then to re-score them using various language models to output the best recognition hypothesis. At present, there exist several open-source and freely available ASR toolkits, web-based tools and engines, which can be adopted by technology developers to any target language using available training data, such as HTK,[15] Julius,[16] Sphinx,[17] RLAT,[18] RASR,[19] KALDI,[20] and YAST,[21] etc.

[15] http://htk.eng.cam.ac.uk.
[16] http://julius.sourceforge.jp/en_index.php.
[17] http://cmusphinx.sourceforge.net.
[18] http://csl.ira.uka.de/rlat-dev.
[19] http://www-i6.informatik.rwth-aachen.de/rwth-asr/.
[20] http://kaldi.sourceforge.net/.
[21] http://pi.imag.fr/xwiki/bin/download/PUBLICATIONS/WebHome/YAST.zip

3.2. Collecting data for UR languages

As mentioned in Section 3.1, the use of statistical modeling motivates the need for (many) data in order to build acoustic, pronunciation and language models. However, for most under-resourced languages, there are no existing corpora that can be used for the development of ASR systems. Hence, data collection is generally an integral part of ASR development in these languages. If we focus on speech data collection, various approaches to data collection for under-resourced languages have been adopted in practice; we distinguish between those that employ existing audio resources, and those that involve the recording of speech as part of the collection process.

In the former category, recordings of radio broadcasts, parliamentary speeches, or similar sources serve as starting point for corpus creation, and the main challenge is to either edit and transcribe the recordings so that they are useful for ASR processes or to leverage off active or unsupervised training methods. Manual transcription is complicated by the common shortage of suitable language practitioners in under-resourced languages; also, many languages do not have well-standardized writing systems (or no writing system at all), in which case the development of suitable corpus-specific standards is a substantial additional burden. Crowd-sourcing approaches to transcription have been used with some success (Parent and Eskenazi, 2010), however, the number of under-resourced languages for which sufficiently many workers are readily available is rather limited and can be very different from one language to another (Gelas et al., 2011). A further complication is that existing sources typically do not have a sufficiently diverse set of speakers for the purposes of ASR. While a typical "speaker-independent" ASR corpus requires at least 50 different speakers (Barnard et al., 2009), radio broadcasts or recordings of lectures may be dominated by a dozen or fewer speakers.

When a corpus is developed from scratch, the transcription task can be simplified significantly, since prompted material can be employed. This benefit must be weighed against the additional burden of soliciting and recording

speakers. In this case, data collection typically starts with the collection of a text corpus (which, again, is only possible if a suitably standardized writing system exists). From this corpus, a collection of prompts are extracted, and presented to selected speakers of the target language for recording. Although verification is still necessary to ensure that speakers did, in fact, say the desired words, automated methods have proven to be quite successful and efficient for this purpose (Davel et al., 2011): an ASR system is bootstrapped from the raw corpus, assuming all prompts were recorded correctly, and this system is used to iteratively identify misspoken utterances and improve the accuracy of the ASR system. For the recording process itself, menu-driven telephone services (also known as Interactive Voice Response services) have often been employed (Muthusamy and Cole, 1992). Instruction sheets containing prompts are distributed to selected speakers of the target language; these speakers call a toll-free number and are guided to record those prompts in order. Alternatively, recordings can be obtained during face-to-face recording sessions (using a tape recorder or personal computer) (Schultz, 2002), such an approach typically benefits from the fact that a field worker can provide personal instructions, but logistical challenges may arise from the fact that all participants have to use one recording device (or perhaps a small number of available devices) in sequence. The widespread availability of smartphones has recently prompted several groups to develop smartphone applications (Hughes et al., 2010; De Vries et al., 2011, 2013) that provide the best of both worlds: personal contact and instruction by a field worker is possible. In that case, the field worker can manage several phones simultaneously, thus enabling the collection of speech from several speakers in a relatively short time.

Of course, spontaneous, rather than prompted, speech can also be collected using any of these platforms (Godfrey et al., 1992). However, such corpora of spontaneous speech are generally less useful as a starting point for ASR development in an under-resourced language: because of resource constraints, relatively small corpora are typically created, and the clearer enunciation of prompted text is relatively more important for such corpora. The difficulties inherent in transcribing spontaneous speech in under-resourced languages mentioned above also favor a prompted approach.

3.3. Feature processing

In the last few years, Neural Networks showed large potential to improve ASR performance. For example, multilayer perceptrons (MLP) were introduced to feature extraction, where the values of the output layer (Tandem features) (Hermansky et al., 2000) or of the hidden layer (Bottle-Neck features) (Grezl et al., 2007) are used in the preprocessing step instead of the traditional MFCC features. In many setups and experimental results, MLP features proved to be of high discriminative power, to be very robust against speaker and environmental variations, and to be somewhat language independent. In the context of ASR for under-resourced languages, those features allow developers to build speech processing systems with small amounts of data and to share speech data of multiple languages to more efficiently bootstrap systems in yet unseen languages.

Several studies showed that features extracted from an MLP which was trained with one or multiple languages can be applied to other languages (Stolcke et al., 2006; Toth et al., 2008; Plahl et al., 2011). Thomas et al. (2012a,b) and Vesely et al. (2012) demonstrated how to use data from multiple languages to extract features for an under-resourced language and, hence, improve ASR performance. They used a data-driven approach in which no prior knowledge about the phone set of the target languages was required. In Vu et al. (2012a,b), the authors presented experiments on using a multilingual MLP for initializing an MLP for under-resourced languages based on IPA phone mapping. The approach showed a substantial improvement in terms of ASR performance and also proved to be robust against transcription errors in the training data (Vu et al., 2012b).

3.4. Acoustic modeling

As mentioned above, it is often difficult to obtain transcriptions of speech in under-resourced languages. Hence, unsupervised or lightly-supervised approaches are particularly attractive in this context. Cetin proposed unsupervised adaptation methods to develop an isolated word recognizer for Tamil (Cetin, 2008), similar and extended approaches have been proposed for Polish (Loof et al., 2009) and for Vietnamese (Vu et al., 2011). Hence, in a scenario where some prior information of the target language is available, such as the pronunciation dictionary, the language model, and the language identification of the untranscribed data, those approaches are very useful to save time and costs by building an ASR system for a yet unsupported language. For instance, the authors in Vu et al. (2011) showed that it is possible even if the source languages and the target language are not related. They used several ASR systems for different languages to decode the audio data of the target language in parallel to compute a confidence score called "Multilingual A-stabil" (Vu et al., 2010). Afterwards, all the words which are voted by at least two different languages are selected to adapt the acoustic model of the target language. In their framework, MAP adaptation was applied iteratively to increase the amount of training data and to improve the automatic transcription quality. In all these developments, transcribed data from well-resourced languages are used to develop initial systems, and untranscribed speech data from the target language (possibly in conjunction with a small amount of transcribed speech) is shown to be sufficient to train usable ASR systems. Interestingly, the relatedness of source and target language is generally not found to be an important

variable: even quite dissimilar languages are found to perform well in this regard.

State-of-the-art ASR systems in well-resourced languages typically employ context-dependent Hidden Markov Models to model the phonemes of a language and the same approach is also commonly used for under-resourced languages. Again, the under-resourced context introduces a number of novel challenges and opportunities. For instance, the definition of an appropriate phoneme set to model is often a non-trivial task: even when such sets have been defined in a language, they often do not have strong empirical foundations (Wissing and Barnard, 2008). Also, putative phonemes such as affricates, diphthongs and click sounds may profitably be modeled as either single units or sequences, and allophones which are acoustically too distinct may be modeled separately. For all these issues, some guidance may be available from choices that have been made in related languages, but some empirical investigation is often required. When a related well-resourced "source" language is available, it may be possible to use data from that language in developing acoustic models for an under-resourced target language. Various approaches have been employed, ranging from pooling data across languages (van Heerden et al., 2010), through bootstrapping from source-model alignments (Schultz and Waibel, 2001; Le and Besacier, 2009), to phone mapping for recognition with the source models (Chan et al., 2012), possibly after some maximum a posteriori (MAP) adaptation with target-language data. Clear guidelines on the best way to perform such cross-lingual sharing, and the amount of benefit that can be expected for different quantities of source and target data, have yet to emerge.

A number of authors have suggested that models other than the standard context-dependent Hidden Markov Models of phonemes are appropriate for under-resourced languages. For example, in exemplar-based speech recognition (see for instance (Gemmeke, 2011)), the representations of acoustic units (words, phonemes) are expressed as vectors of weighted examples. Such methods, with low number of parameters, appear to be particularly interesting if little data is available for training. A less radical departure from the standard model uses Hidden Markov Models to model syllables rather than phonemes (Tachbelie et al., 2012, 2013), in this case, the reduction in model parameters results from the fact that context dependencies are generally less important for syllable models. Siniscalchi et al. (2013) proposes to describe any spoken language with a common set of fundamental units that can be defined "universally" across all spoken languages. Speech attributes, such as manner and place of articulation (similar to those proposed by Stuker et al. (2003)), are chosen to form this unit inventory and used to build a set of language-universal attribute models derived from IPA (Stüker et al., 2003) or with data-driven modeling techniques. The latter work proposed by Siniscalchi et al. (2013) is well suited for deep neural network architectures for ASR (Yu et al., 2012).

3.5. Lexical modeling

3.5.1. Grapheme-based approaches

Regarding the creation of pronunciation dictionaries, grapheme-based approaches were presented for many languages, such as Thai (Charoenpornsawat et al., 2006; Stüker, 2008), Amharic (Gizaw, 2008), Vietnamese (Le and Besacier, 2009) and even for multiple languages (Killer et al., 2003; Kanthak and Ney, 2003). In grapheme-based modeling, each word in the pronunciation dictionary is simply decomposed into its graphemes; these graphemes being the basic units of the acoustic model. Such systems give decent results, particularly for those languages with a close grapheme-to-phoneme relationship.

3.5.2. Bootstrapping G2P using MT approaches

Other approaches of converting graphemes to phonemes use statistical machine translation principles (Laurent et al., 2009; Karanasou and Lamel, 2010). Here, graphemes are regarded as "words" in the source language and the phonemes as "words" in the target language. A "machine translation" system is trained based on an initial phonetic dictionary and afterwards this system is applied to convert any word to its phonetic form. Such an approach was, for example, proposed for Romanian language in Cucu et al. (2011).

3.5.3. Use of the Web

Ghoshal et al. (2009), Schlippe et al. (2010, 2013) describe automatic methods to produce pronunciation dictionaries using word-pronunciation pairs found in the World Wide Web. Since Wiktionary (a wiki-based open content dictionary) contains phonetic notations written in the International Phonetic Alphabet (IPA, 1999), (Schlippe et al., 2010) developed a system which automatically extracts phonetic notations in IPA from Wiktionary. The authors reported results for the four languages English, French, German, and Spanish concerning quantity and quality checks. The quantity checks with lists of international cities and countries demonstrated that even proper names, for which pronunciations might not be found in the phonetic system of a language can be retrieved from Wiktionary along with their phonetic notations. However, this appeared to strongly depend on the quantity and quality of the data found on Wiktionary. Unfortunately, the majority of the languages in the world are not covered yet in Wiktionary. In Schlippe et al. (2012a,b), the G2P model generation for Indo-European languages was investigated with word-pronunciation pairs from 6 Wiktionary editions and 10 GlobalPhone dictionaries. Using pronunciations exclusively generated from Wiktionary, G2P models for ASR training and decoding resulted in reasonable performance degradations given the cost and time efficient generation process. Schlippe et al. (2012b) propose fully automatic methods to detect, remove, and substitute incon-

sistent or flawed word-pronunciation entries from the World Wide Web and showed quality improvements.

3.6. Language modeling

Statistical language models provide an estimate of the probability of a word sequence. One of the most efficient statistical language modeling schemes is based on word *n*-grams (bigrams, trigrams, and more) that estimate the probability of any word sequence in some text. The probabilities in *n*-gram language models are commonly determined by means of maximum likelihood estimation. This makes the probability distribution dependent on the available training data. Thus, to ensure statistical significance, large training data are required in statistical language modeling.

3.6.1. Word decomposition and use of syntactic information

For some morphologically-rich languages, it is efficient to decompose words into sub-lexical units (morphemes or rather morphs as realizations of morphemes in text data) and apply them as tokens in the vocabulary and LM. Such technique allows reducing the recognition vocabulary and provides better lexicon coverage resulting in a smaller amount of out-of-vocabulary (OOV) words. However, it makes also some additional challenges at speech decoding, including a high phonetic ambiguity of sub-word units, specific grapheme-to-phoneme conversion with multiple transcriptions, necessity to compose whole-words from recognized particles, as well as higher order *n*-grams (5- to 10-grams) are required to capture grammatical dependencies. Morpheme-based models were successfully applied for some (in particular, agglutinative and inflective) languages, such as Finnish (Creutz et al., 2007), Turkish (Sak et al., 2010; Arisoy et al., 2006; Carki et al., 2000), Estonian (Kurimo et al., 2006a,b), Hungarian (Tarjan and Mihajlik, 2010; Szarvas and Furui, 2003), Czech (Oparin et al., 2008), Slovenian (Rotovnik et al., 2007), Russian (Whittaker, 2000; Ronzhin and Karpov, 2007), and even German (Adda-Decker, 2003). Particle-based LMs were also successfully realized for morphologically-rich non-European languages such as Arabic (Vergyri et al., 2004; Sarikaya et al., 2007), Amharic (Pellegrini and Lamel, 2009; Tachbelie et al., 2012), Korean (Kiecza et al., 1999; Le and Rim, 2009), and Uyghur (Ablimit et al., 2010) (both morphemic and syllabic LMs), etc. In practice, decomposition of word-forms into morphs can be performed by two different approaches: grammatical (knowledge-based) methods and statistical (unsupervised) methods based on statistical analysis of a large text corpus (Kurimo et al., 2006b). The advantage of grammatical methods is that they allow obtaining a genuine decomposition of the word-forms into lexical morphemes. The feature of the statistical methods is that they rely on a text analysis only and do not use any additional linguistic knowledge, so texts written in any language can be processed; however, words may be divided into pseudo-morpheme units by these methods. There are some widely used software for unsupervised word decomposition, for instance Morfessor (Creutz and Lagus, 2005) that was originally developed for Finnish.[22]

As far as (under-resourced) language modeling is concerned, text data sparseness is a very challenging issue. This problem was addressed in several studies: for instance for two African languages: Somali (Abdillahi et al., 2006) and Amharic (Pellegrini and Lamel, 2006; Tachbelie et al., 2013) and one Eastern European language: Hungarian (Mihajlik et al., 2007). These papers proposed word decomposition algorithms for language modeling in order to reduce the vocabulary size. In Pellegrini and Lamel (2008), interesting experiments to measure the relative importance of text training data for ASR in less-resourced languages are also presented; in the same paper, minimum requirements on the data quantities needed to build an ASR system are suggested.

Some under-resourced languages, for instance, Slavic languages (Ukrainian, Russian, Belarusian, Czech, Slovak, Slovene, etc.), are characterized by practically free order of words in sentences in contrast to many fixed word-order languages like English or German. Syntactic and semantic information is crucial for determining correct order of words and sentence structure. Standard statistical language models are not so efficient for these languages because high order *n*-grams (trigrams and more) have a high perplexity and a low *n*-gram hit rate, so huge corpora are needed to estimate probabilities for these models. There are some recent works that suggest taking into account syntactical information and long distance dependencies between words in sentences simultaneously with statistical language modeling, for instance, structured language models (Chelba and Jelinek, 2000) and some enhanced *n*-gram models (Kanejiya et al., 2003; Rastrow et al., 2012; Kuo et al., 2009; Kipyatkova et al., 2012; Karpov et al., 2013). Also, some syntactical information obtained by automatic text parsers can be used to capture and model grammatical dependencies contained in sentences (Lopatková et al., 2005; Charniak et al., 2003; Huet et al., 2010), resulting in better recognition accuracy.

3.6.2. Web or translation-based text data collection

The collection of textual data in a given language (and for a given domain) is also a hot topic that can be addressed using the Web as a corpus (Le et al., 2003; Cai, 2008) or using machine translation systems to port text corpora from one language to another (Nakajima et al., 2002; Jensson, 2008; Suenderman and Liscombe, 2009; Cucu et al., 2012). However, one faces specific problems, when developing language models for some under-resourced languages. For instance, languages like Romanian or Turkish make intensive use of diacritics. Even though for a human reader the meaning of a text without diacritics is most of the times obvious (given the surround-

[22] http://www.cis.hut.fi/projects/morpho/.

ing context), machine diacritics restoration is not a trivial task and it is important in some contexts. For instance, for several languages that use diacritics, text corpora which can be acquired over the web come without diacritics. The output of an ASR system lacking diacritics could be ambiguous or even incomprehensible. Therefore, an automatic diacritics restoration system is mandatory for these languages (see Cucu et al. (2013) for instance). Other technical issues are the need for normalization (numbers, acronyms, abbreviations, etc.) as well as the use of language identification as a pre-processing to filter out web pages in a different language. Spelling errors and inconsistencies in the writing system are also important problems to be dealt with in under-resourced languages context.

3.6.3. Word segmentation issues

The writing systems of some languages like Chinese, Vietnamese, Khmer, and Thai lack word separators completely or use them inconsistently. The definition of word units is crucial for ASR, as the dictionary and the language model rely on it. The segmentation into word units or "word identification" is not a trivial task even for languages that separate words by a special character (a white-space in general). For languages, which have a writing system without obvious separation between words, the *n*-grams of words are usually estimated from a text corpus segmented into words employing automatic methods. Automatic segmentation of text is not a trivial task and introduces errors due to the ambiguities in natural language and the presence of out of vocabulary words in the text. A possible alternative is to calculate the probabilities from logographic characters (e.g. Kanji in Japanese or Hanja in Korean) like in Denoual and Lepage (2006).

3.7. Evaluating ASR performance

Word Error Rate (WER) is an intuitive and adequate measure for word-oriented analytical languages with quite simple morphology; however, some languages are morpheme-based while some others are syllable-based. Moreover, as said earlier, some languages (e.g. Thai, Vietnamese) have not obvious separators between orthographic words. So, these languages can synthesize quite long meaningful word-forms from a number of sub-word units. For example, in many agglutinative languages like Estonian or Finnish, word-forms can be composed of a root (stem) preceded or followed by up to dozen grammatical affixes and such ending is usually pronounced not as clearly as the beginning part that results in acoustic and phonetic ambiguity and higher WER. For ASR of morphologically-rich languages, some more adequate metrics can be applied: Letter/Character Error Rate (LER or CER) (Kurimo et al., 2006a,b), Phone Error Rate (PER), Syllable Error Rate (SylER) (Huang et al., 2000) or Morpheme Error Rate (Ablimit et al., 2010). There exist also some other measures, such as Inflectional Word Error Rate (IWER) (Bhanuprasad and Svenson, 2008; Karpov et al., 2011), Speaker Attributed Word Error Rate (NIST, 2009), Weighted Word Error Rate (WWER) (Nanjo and Kawahara, 2005), etc.

4. Applications and Tools for U-ASR

4.1. Voice search in three South African languages

South Africa is a highly diverse country, with wide social disparities and eleven official languages. Technology projects that address social issues while also bridging language barriers have therefore achieved substantial attention in South African in recent years (Barnard et al., 2010), and substantial progress has been made in developing speech resources and systems that encompass all eleven languages. A highly visible (and commercially relevant) result of this activity was the development of applications that perform Web searches based on spoken queries in three South African languages, namely isiZulu, South African English and Afrikaans (Barnard et al., 2010). Using several of the techniques described above, resources were collected and ASR systems were developed using tools and infrastructure provided by Google; these systems were found to be somewhat less accurate than state-of-the-art systems in American English, but of sufficient quality to be released commercially. Both the Afrikaans and the South African English have attracted active user populations; in isiZulu, however, the amount of information available on the Web is too limited to support an active user base.

4.2. Interactive voice forum for farmers in rural India

This project (called Avaaj Otalo) was designed in the summer of 2008 as a joint project between a Non Governmental Organization in India and IBM India Research Laboratory. A voice message forum was proposed to farmers in India (who often have limited formal education) to provide interactive on-demand access to agricultural knowledge. Voice content was accessed using low-cost mobile phones, which are being rapidly adopted by rural communities around the world. The most popular feature of the project was a forum for asking questions and browsing others' questions and responses on a range of agricultural topics (check weather reports for help them decide when to fertilize crops, know when doctors are coming into town, find the best prices for their crops or merchandise, etc.). As far as ASR is concerned, user inputs were forwarded to the speech recognition engine, IBM's Websphere Voice Server (WVS). Since WVS is a large vocabulary, continuous speech recognizer trained on American English, it had to be adapted to Gujarati language considered in the project (spoken by $\sim$50 M persons in India). For this, Gujarati speech commands were converted using the American English phoneme set. With this approach, a speech recognition accuracy of 94% in a largely quiet, indoor setting was observed (see Patel et al. (2009) for more

details). However, in terms of usability, it was shown later in Patel et al. (2010) that for simple menu-based navigation, users preferred numeric input over speech.[23]

4.3. The PI project

The PI project (funded by French ANR – Agence Nationale de la Recherche) was fully dedicated to automatic speech recognition for under-resourced languages, especially languages from Vietnam, Laos and Cambodia. From an operational point of view, this project aimed at providing tools for ASR development in under-resourced languages (all project deliverables – reports or software – can be downloaded from the project website[24]). Another result of the PI project was a strong contribution to the structuring of the scientific community around the topic "processing under-resourced languages" (see Section 5.4 that summarizes events organized or co-organized by the PI project participants).

4.4. The Rapid Language Adaptation toolkit (RLAT)

The project SPICE (NSF, 2004–2008) performed at the Language Technologies Institute at Carnegie Mellon and the Rapid Language Adaptation project at the Cognitive Systems Lab (CSL) aimed at bridging the gap between the language and technology expertise. For this purpose RLAT[25] provides innovative methods and interactive web-based tools to enable users to develop speech processing models in any language, to collect appropriate speech and text data to build these models, as well as to evaluate the results allowing for iterative improvements. The toolkit significantly reduces the amount of time and effort involved in building speech processing systems for unsupported languages. In particular, the toolkit allows the user to (1) design databases for new languages at low cost by enabling users to record appropriate speech data along with transcriptions, (2) to continuously harvest, normalize, and process massive amounts of text data from the web, (3) to select appropriate phone sets for new languages efficiently, (4) to create vocabulary lists, (5) to automatically generate pronunciation dictionaries, (6) to apply these resources by developing acoustic and language models for speech recognition, (7) to develop models for text-to-speech synthesis, and (8) to finally integrate the built components into an application and evaluate the results using online speech recognition and synthesis in a talk-back function (Schultz et al., 2007). RLAT and SPICE are a freely available online services which provides an interface to the web-based tools and has been designed to accommodate all potential users, ranging from novices to experts. The tools are regularly used for training and teaching purposes at two universities (KIT and CMU). Results indicate that it is feasible to build end-to-end speech processing systems in various languages (more than 15) for small domains within the framework of a six-week hands-on lab course.

5. The future of U-ASR

5.1. Endangered languages

As already said, language diversity is fragile as some languages are threatened or in real danger of extinction. With such a perspective, revitalization and documentation programs are emerging.[26] So, while there is commercial interest in enabling the ~300 most widely spoken languages in the digital domain (if digital technologies work for this group of languages that represents 95% of humanity), there are other reasons to work on the other ~6500 languages that are not of commercial interest: to provide access to information, to provide a critical new domain of use for endangered languages, for better linguistic knowledge of them, for response in a crisis ("surge languages"), etc. We are convinced that using automatic speech recognition technologies would be particularly useful for computer assisted language learning of the endangered languages. In addition, the development of tools for field linguists (automatic annotation tools, forced alignment and segmentation, etc.) seems important for revitalizing or at least for documenting endangered languages. The idea here is to evaluate the analysis capabilities of existing automatic speech processing systems to investigate phonetic characteristics of languages. For instance, Gelas et al. (2010) showed the relevance of multilingual acoustic models to study, at a large scale, particular phenomena of rare languages.

5.2. Non written languages

As said in Section 1, if we want to address all languages in the world, we have to prepare for encountering many languages without a writing system. In such a context, it is interesting to address the problem of automatically exploring non written languages for which no ASR or MT systems have been created so far. One can imagine a particular scenario where a human translator is available and where engineers try to exploit the translations of this human interpreter (utterances in the non-written target language), in order to gather the material needed for training ASR and translation systems. If the language is unwritten, one can only work with a phonetic transcription of that language (or with the signal itself). Such a transcription can be obtained manually by skilled phoneticians or using multilingual acoustic decoders (as seen in Section 3.5). In

[23] http://www.watblog.com/2012/01/16/speech-driven-web-service-for-indian-farmers-launched-by-indian-govt/.
[24] http://pi.imag.fr/xwiki/bin/view/PUBLICATIONS/.
[25] Rapid Language Adaptation Toolkit (RLAT) http://csl.anthropomatik.kit.edu/rlat.php.
[26] See for instance "Sorosoro" program funded by the Chirac Foundation http://www.sorosoro.org/.

Besacier et al. (2006) and Stüker et al. (2009) feasibility of automatically learning word units (as well as their pronunciation) without any supervision, in the unknown language, was examined,. This was done by unsupervised aggregation of phonetic strings (to form words) from a continuous flow of phonemes (or from a signal). In the scenario where a human translator produces utterances in the (unwritten) target language from English prompts, adding the English source to help the word discovery process was shown to be efficient. An overview of the approaches for "human translations guided language discovery for ASR" can be found in Stüker et al. (2009). Stahlberg et al. (2012) proposed Model3P, an extended version of the alignment model IBM Model3, to improve the aggregation of the phoneme strings. In Stahlberg et al. (2013) phonetic transcriptions of target language words using Model3P were deduced and then introduced in the pronunciation dictionary. Analyzing 14 translations in 9 languages to build a dictionary in an unknown target language showed that the quality of the resulting dictionary is better in case of close vocabulary sizes between source and target language, shorter sentences, more word repetitions, and formal equivalent translations.

5.3. Tasks beyond U-ASR

More and more research works are published on under-resourced languages issues for HLT tasks beyond ASR. For instance, text-to-speech systems have been developed for several languages, and are the topic of a couple of papers in the current Special Issue (van Niekerk and Barnard, 2013; Ekpenyong et al., 2013). Also, machine translation for under-resourced language pairs is becoming increasingly popular. Good examples are Do et al. (2010) for Vietnamese–French translation and (Gebreegziabher and Besacier, 2012) for Amharic–English MT. The problem with machine translation is that for many language pairs, cross-language resources are scarce. In addition to the case of under-resourced languages that have scarce resources by themselves, it is also an important issue for pairs of well-resourced languages that have few parallel resources (because of their cultural, historical and/or geographical disconnection, for instance, Spanish–Chinese language pair). This is also the case for single languages for which new communication trends and styles do not have available cross-language resources between the main formal language and its informal versions (as chat speaking style, communications, and formal languages). Recently, an LREC 2012 workshop[27] was dedicated to these issues and introduced the concept of disconnected languages and styles.

[27] http://www-lium.univ-lemans.fr/credislas2012/.

5.4. Organizing the research community on U-ASR

The authors of this paper have already initiated some networking activity around the topic of under-resourced languages, as illustrated below with a list of events (chronological order) organized or co-organized by one or several authors of this paper:

- Workshop SLTU (Spoken Language Technologies for Under-Resourced Languages) 2008[28]
- Workshop SLTU 2010[29]
- African HLT 2010 in Djibouti[30]
- Tutorial on Rapid Language Adaptation Tools & Technologies at ICASSP 2008
- Tutorial on Rapid Language Adaptation Tools & Technologies at Interspeech 2010
- Special Session on Under-Resourced Languages at Interspeech 2011[31]
- Workshop SLTU 2012[32]
- Workshop on African Language Processing during JEP-TALN 2012 (in French)[33]
- Organization of a tutorial during the 3L Summer School on Endangered Languages in 2012[34]

One important result of this networking activity is a bi-annual workshop called Spoken Language Technologies for Under-resourced languages (SLTU) that will have its 4th edition in 2014.[35] Some scientific organizations are also very active on this topic: research on HLT for languages of East Africa is well structured through AfLaT (African Language Technology) organization,[36] however AfLat has a lower impact in countries (notably in Western Africa) where, in the scientific community, French is preferred to English. The International Speech Communication Association (ISCA) has a special interest group called SALTMIL.[37] (Speech and Language Technologies for Minority Languages) but, as said in Section 3.1, minority languages are not the same as under-resourced languages which can be official and/or national languages of their country and spoken by a very large population. So, the need for an international organization for processing under-resourced languages remains. The publication of this special issue on Processing Under-Resourced Languages is an important step into this direction. The development of a special interest group (SIG) on this topic at ISCA is another step. Last but not least, ambitious research projects, funded by

[28] http://www.mica.edu.vn/sltu/.
[29] http://www.mica.edu.vn/sltu-2010/.
[30] http://www.lanation.dj/news/2010/ln14/national8.htm.
[31] http://www.interspeech2011.org/specialsessions/ss-7.html.
[32] http://www.mica.edu.vn/sltu2012/.
[33] http://www.jeptaln2012.org/actes/TALAF2012/index.html.
[34] http://www.ddl.ish-lyon.cnrs.fr/colloques/31_2012/index.asp?Langues=EN&Page=Programme.
[35] http://www.mica.edu.vn/sltu2014/.
[36] http://aflat.org.
[37] http://ixa2.si.ehu.es/saltmil/.

international organizations such as EU, ASEAN or UNESCO, would help to gather main research and industrial actors interested in processing under-resourced languages at a large scale.

6. Conclusion

Our survey and the papers in this Special Issue demonstrate that speech processing for under-resourced languages is an active field of research, which has experienced significant progress during the past decade. The current review has focused on speech recognition, since that is the area which has been the most significant focus of research for these languages; however, it should be clear that many of the issues and approaches apply to speech technology in general. Although much of the recent progress has been the result of the technical developments summarized in Section 3, it is clear that organizational developments will be required to address many of the pertinent issues. In particular, progress with the smaller languages and those with extremely limited resources (such as the language mentioned in Section 5) will most likely rely on significant resource sharing; however, such sharing will benefit greatly from organizations and facilities that make it easy for researchers and technologists to access available resources in a wide range of languages. It is our hope that the current wave of interest in under-resourced languages will stimulate cooperation along these lines, along with continuing scientific research to support such languages – and, ultimately, their speakers.

References

Abdillahi, N., Nocera, P., Bonastre, J.-F., 2006. Automatic transcription of Somali language. In: ICSLP'06, Pittsburgh, PA, USA, pp. 289–292.

Ablimit, M., Neubig, G., Mimura, M., Mori, S., Kawahara, T., Hamdulla, A., 2010. Uyghur Morpheme-based language models and ASR. In: Proc. IEEE 10th International Conference on Signal Processing (ICSP), Beijing, China, pp. 581–584.

Adda-Decker, M., 2003. A corpus-based decompounding algorithm for German lexical modeling in LVCSR. In: Proc. Eurospeech-2003, Geneva, Switzerland, pp. 257–260.

Arisoy, E., Dutagaci, H., Arslan, L., 2006. A unified language model for large vocabulary continuous speech recognition of Turkish. Signal Processing 86 (10), 2844–2862.

Arisoy, E., Sainath, T.N., Kingsbury, B., Ramabhadran, B., 2012. Deep neural network language models. In: Proc. NAACL-HLT 2012 Workshop, Montreal, Canada, pp. 20–28.

Barnard, E., Davel, M., van Heerden, C., 2009. ASR corpus design for resource-scarce languages. In: Proc. Interspeech, pp. 2847–2850.

Barnard, E., Davel, M., van Huyssteen, G.B., 2010. Speech technology for information access: a South African case study. In: Proceedings of the AAAI Spring Symposium on Artificial Intelligence for Development (AI-D), Palo Alto, California, March 2010, pp. 8–13.

Barnett, J., Corrada, A., Gao, G., Gillik, L., Ito, Y., Lowe, S., Manganaro, L., Peskin, B., 1996. Multilingual speech recognition at Dragon systems. In: Proc. ICSLP, Philadelphia, pp. 2191–2194.

Berment, V., 2004. Méthodes pour informatiser des langues et des groupes de langues peu dotées. Ph.D. Thesis, J. Fourier University – Grenoble I, May 2004.

Besacier, L., Zhou, B., Gao, Y., 2006. Towards speech translation of non written languages. In: IEEE/ACL SLT 2006. Aruba, December 2006.

Bhanuprasad, K., Svenson, M., 2008. Errgrams – a way to improving ASR for highly inflective Dravidian languages. In: Proc. 3rd International Joint Conf. on Natural Language Processing IJCNLP'08, India, pp. 805–810.

Billa, J., Ma, K., McDonough, J., Zavaliagkos, G., Miller, D.R., Ross, K.N., El-Jaroudi, A., 1997. Multilingual speech recognition: the 1996 Byblos Callhome system. In: Proc. Eurospeech-1997, Rhodes, Greece, pp. 363–366.

Cai, J., 2008. Transcribing southern min speech corpora with a web-based language learning system. In: SLTU'08, Hanoi, Vietnam.

Carki, K., Geutner, P., Schultz, T., 2000. Turkish LVCSR: towards better speech recognition for agglutinative languages. In: IEEE ICASSP.

Cetin, O., 2008. Unsupervised adaptive speech technology for limited resource languages: a case study for Tamil. In: SLTU'08, Hanoi, Vietnam.

Chan, H.Y., Rosenfeld, R. 2012. Discriminative pronunciation learning for speech recognition for resource scarce languages. In: Proceedings of the 2nd ACM Symposium on Computing for Development. Article No. 12.

Charniak, E., Knight, K., Yamada, K., 2003. Syntax-based language models for machine translation. In: Proc. IX MT Summit, New Orleans, USA, pp. 40–46.

Charoenpornsawat, P., Hewavitharana, S., Schultz, T., 2006. Thai grapheme-based speech recognition. In: Human Language Technology Conference (HLT).

Chelba, C., Jelinek, F., 2000. Structured language model. Computer Speech and Language 10, 283–332.

Cohen, P., Dharanipragada, S., Gros, J., Monkowski, M., Neti, C., Roukos, S., Ward, T., 1997. Towards a universal speech recognizer for multiple languages. In: Proc. Automatic Speech Recognition and Understanding (ASRU), St. Barbara CA, pp. 591–598.

Constantinescu, A., Chollet, G., 1997. On cross-language experiments and data-driven units for ALISP. In: Proc. Automatic Speech Recognition and Understanding (ASRU), St. Barbara CA, pp. 606–613.

Creutz, M., Lagus, K., 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. Computer and Information Science, Report A81, Helsinki University of Technology, Finland.

Creutz, M., Hirsimaki, T., Kurimo, M., Puurula, A., Pylkkonen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saraclar, M., Stolcke, A., 2007. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. ACM Transactions on Speech and Language Processing 5 (1). Article No. 3.

Crystal, D., 2000. Language Death. Cambridge CUP.

Cucu, H., Besacier, L., Burileanu, C., Buzo, A., 2011. Investigating the role of machine translated text in ASR domain adaptation: unsupervised and semi-supervised methods. In: Proc. ASRU 2011, Hawaii, USA.

Cucu, H., Besacier, L., Burileanu, C., Buzo, A., 2012. ASR domain adaptation methods for low-resourced languages: application to Romanian language. In: EUSIPCO'2012, Bucarest, Romania.

Cucu, H., Buzo, A., Besacier, L., Burileanu, C., 2013. SMT-based ASR domain adaptation methods for under- resourced languages: application to Romanian. Speech Communication. http://dx.doi.org/10.1016/j.specom.2013.05.003.

Davel, M.H., van Heerden, C., Kleynhans, N., Barnard, E., 2011. Efficient harvesting of Internet audio for resource-scarce ASR. In: Proc. Interspeech, pp. 3153–3156.

De Vries, N.J., Badenhorst, J., Davel, M.H., Barnard, E., De Waal, A., 2011. Woefzela-an open-source platform for ASR data collection in the developing world. In: Proc. Interspeech, pp. 3177–3180.

De Vries, N.J., Davel, M.H., Badenhorst, J., Basson, W.D., de Wet, F., Barnard, E., De Waal, A., 2013. A smartphone-based ASR data collection tool for under-resourced languages, Speech Communication. http://dx.doi.org/10.1016/j.specom.2013.07.001.

Denoual, E., Lepage, Y., 2006. The character as an appropriate unit of processing for non-segmenting languages. In: NLP Annual Meeting, Tokyo, Japan, pp. 731–734.

Do, T., Besacier, L., Castelli, E., 2010. Unsupervised SMT for a low-resourced language pair. In: Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU), Penang, Malaysia.

Dugast, C., Aubert, X., Kneser, R., 1995. The Philips large-vocabulary recognition system for American English, French, and German. In: Proc. Eurospeech, Madrid, pp. 197–200.

Ekpenyong, M., Urua, E.-A., Watts, O., King, S., Yamagishi, J., 2013. Statistical parametric speech synthesis for Ibibio, Speech Communication. http://dx.doi.org/10.1016/j.specom.2013.02.003.

Ganapathiraju, A., Hamaker, J., Picone, J., 2000. Hybrid SVM/HMM architectures for speech recognition. In: Proceedings of Speech Transcription Workshop, pp. 504–507.

Gebreegziabher, M., Besacier, L., 2012. English-Amharic statistical machine translation. In: SLTU – Workshop on Spoken Language Technologies for Under-Resourced Languages, Cape-Town, South Africa.

Gelas, H., Besacier, L., Rossato, S., Pellegrino, F., 2010. Using automatic speech recognition for phonological purposes: study of vowel length in Punu (Bantu B40). In: Laphon 12, New Mexico (US), July 2010.

Gelas, H., Teferra Abate, S., Besacier, L., Pellegrino, F., 2011. Quality assessment of crowdsourcing transcriptions for African languages. In: Interspeech 2011 Florence, Italy, 28–31 August 2011.

Gemmeke, J.F., Van hamme, H., 2011. A hierarchical exemplar-based sparse model of speech with an application to ASR. IEEE ASRU 2011, HI, USA.

Ghoshal, A., Jansche, M., Khudanpur, S., Riley, M., Ulinski, M., 2009. Web-derived pronunciations. In: IEEE ICASSP.

Gizaw, S., 2008. Multiple pronunciation model for Amharic speech recognition system. In: SLTU 2008, Hanoi, Vietnam.

Glass, J., Flammia, G., Goodine, D., Phillips, M., Polifroni, J., Sakai, S., Seneff, S., Zue, V., 1995. Multi-lingual spoken language understanding in the MIT voyager system. Speech Communication 17, 1–18.

Godfrey, J.J., Holliman, E.C., McDaniel, J., 1992. SWITCHBOARD: telephone speech corpus for research and development. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 517–520.

Gokcen, S., Gokcen, J., 1997. A multilingual phoneme and model set: towards a universal base for automatic speech recognition. In: Proc. Automatic Speech Recognition and Understanding (ASRU), St. Barbara CA, pp. 599–603.

Grezl, F., et al., 2007. Probabilistic and bottle-neck features for LVCSR of meetings. In: Proc. ICASSP, USA.

Hermansky, H., Wellis, D., Sharma, S., 2000. Tandem connectionist feature extraction for conventional HMM systems. In: Proc. ICASSP, Turkey.

Huang, C., Chang, E., Zhou, J., Lee K.-F., 2000. Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition. In: Proc. INTERSPEECH-2000, Beijing, China, pp. 818–821.

Huet, S., Gravier, G., Sebillot, P., 2010. Morpho-syntactic postprocessing of N-best lists for improved French automatic speech recognition. Computer Speech and Language 24 (4), 663–684.

Hughes, T., Nakajima, K., Ha, L., Moreno, P., LeBeau, M., 2010. Building transcribed speech corpora quickly and cheaply for many languages. In: Proc. Interspeech, Makuhari, Japan, pp. 1914–1917.

IPA, 1999. Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet. Cambridge University Press.

Jensson, A., 2008. Development of a speech recognition system for Icelandic using machine translated text. In: SLTU'08, Hanoi, Vietnam.

Jing, Z., Min, Z., 2010. Speech recognition system based improved DTW algorithm. In: Proc. Int. Conf. on Computer, Mechatronics, Control and, Electronic Engineering CMCE-2010, vol. 5, pp. 320–323.

Kanejiya, D.P., Kumar, A., Prasad, S., 2003. Statistical language modeling using syntactically enhanced LSA. In: Proc. TIFR Workshop on Spoken Language Processing, Mumbai, India, pp. 93–100.

Kanthak, S., Ney, H., 2003. Multilingual acoustic modeling using graphemes. In: Eurospeech-2003, Geneva, Switzerland, pp. 1145–1148.

Karanasou, P., Lamel, L., 2010. Comparing SMT methods for automatic generation of pronunciation variants. In: IceTAL 2010, Reykjavik, Iceland, p. 167.

Karpov, A., Kipyatkova, I., Ronzhin, A., 2011. Very large vocabulary ASR for spoken Russian with syntactic and morphemic analysis. In: Proc. Interspeech'2011, Florence, Italy, pp. 3161–3164.

Karpov, A., Markov, K., Kipyatkova, I., Vazhenina, D., Ronzhin, A., 2013. Large vocabulary Russian speech recognition using syntactico-statistical language modeling. Speech Communication. http://dx.doi.org/10.1016/j.specom.2013.07.004.

Kiecza, D., Schultz, T., Waibel, A., 1999. Data-driven determination of appropriate dictionary units for Korean LVCSR. In: Proceedings of the International Conference on Speech Processing, pp. 323–327.

Killer, M., Stüker, S., Schultz, T., 2003. Grapheme based speech recognition. In: Interspeech.

Kipyatkova, I., Karpov, A., Verkhodanova, V., Zelezny, M., 2012. Analysis of long-distance word dependencies and pronunciation variability at conversational Russian speech recognition. In: Proc. FedCSIS-2012, Wroclav, Poland, pp. 719–725.

Köhler, J., 1998. Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks. In: Proc. ICASSP, Seattle, pp. 417–420.

Krauwer, S., 2003. The basic language resource kit (BLARK) as the first milestone for the language resources roadmap. In: Proceedings of the 2003 International Workshop Speech and Computer SPECOM-2003, Moscow, Russia, pp. 8–15.

Kuo, H.-K.J., Mangu, L., Emami, A., Zitouni, I., Lee, Y.-S., 2009. Syntactic features for Arabic speech recognition. In: Proc. International Workshop ASRU'2009, Merano, Italy, pp. 327–332.

Kurimo, M., Puurula, A., Arisoy, E., Siivola, V., Hirsimaki, T., Pylkkonen, J., Alumae, T., Saraclar, M., 2006. Unlimited vocabulary speech recognition for agglutinative languages. In: Proc. HLT-NAACL, NY, USA.

Kurimo, M., et al., 2006. Unsupervised segmentation of words into morphemes – Morpho Challenge. Application to automatic speech recognition. In: Proc. Interspeech'06, Pittsburgh, PA, USA, pp. 1021–1024.

Lamel, L., Adda-Decker, M., Gauvain, J.L., 1995. Issues in large vocabulary multilingual speech recognition. In: Proc. Eurospeech, Madrid, pp. 185–189.

Laurent, A., Deléglise, P., Meignier, S., 2009. Grapheme to phoneme conversion using an SMT system. In: Interspeech 2009, Brighton, UK, pp. 708–711.

Le, V.-B., Besacier, L., 2009. Automatic speech recognition for under-resourced languages: application to Vietnamese language. IEEE Transactions on Audio, Speech and Language Processing 17(8), 1471–1482.

Le, V.B., Bigi, B., Besacier, L., Castelli, E., 2003. Using the Web for fast language model construction in minority languages. In: Eurospeech'03, Geneva, Switzerland, pp. 3117–3120.

Lee, D.-G., Rim, H.-C., 2009. Probabilistic modeling of Korean morphology. IEEE Transactions on Audio, Speech & Language Processing 17 (5), 945–955.

Loof, J., Gollan, C., Ney, H., 2009. Cross-language bootstrapping for unsupervised acoustic model training: rapid development of a Polish speech recognition system. In: Interspeech 2009. Brighton, UK.

Lopatková, M., Plátek, M., Kuboň, V., 2005. Modeling syntax of free word-order languages: dependency analysis by reduction. In: Proc. TSD'2005, Springer LNAI 3658, Karlovy Vary, Czech Republic, pp. 140–147.

Mihajlik, P., Fegyó, T., Tüske, Z., Ircing, P., 2007. Morpho-graphemic approach for the recognition of spontaneous speech in agglutinative languages – like Hungarian. In: Interspeech'07, Antwerp, Belgium.

Mikolov, T., Karafiat, M., Burget, L., Cernocky, J., Khudanpur, S., 2010. Recurrent neural network based language model. In: Proc. INTERSPEECH-2010, Makuhari, Japan, pp. 1045–1048.

Mohamed, A., Dahl, G.E., Hinton, G., 2012. Acoustic modeling using deep belief networks. IEEE Transactions on Audio, Speech, and Language Processing 20 (1), 14–22.

Muthusamy, Y.K., Cole, R.A., 1992. Automatic segmentation and identification of ten languages using telephone speech. In: Second International Conference on Spoken Language Processing.

Nakajima, H., Yamamoto, H., Watanabe, T., 2002. Language model adaptation with additional text generated by machine translation. In: COLING 2002, vol. 2, Taipei, Taiwan, pp. 716–722.

Nanjo, H., Kawahara, T., 2005. A new ASR evaluation measure and minimum Bayes-risk decoding for open-domain speech understanding. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP-2005, PA, USA, pp. 1053–1056.

The US NIST 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan, 2009.

Oparin, I., Glembek, O., Burget, L., Černocký, J., 2008. Morphological random forests for language modeling of inflectional languages. In: Proc. IEEE Workshop on Spoken Language Technology SLT'08, Goa, India.

Parent, G., Eskenazi, M., 2010. Toward better crowdsourced transcription: transcription of a year of the Let's Go bus information system data. In: Proceedings of IEEE Workshop on Spoken Language Technology, Berkeley, California, December 2010, pp. 312–317.

Patel, N., Agarwal, S., Rajput, N., Nanavati, A., Dave, P., Parikh, T.S., 2009. A comparative study of speech and dialed input voice interfaces in rural India. In: CHI '09: Proceedings of the 27th international conference on Human factors in computing systems. ACM, New York, NY, USA, pp. 51–54.

Patel, N., Chittamuru, D., Jain, A., Dave, P., Parikh, T.S., 2010. Avaaj Otalo: a field study of an interactive voice forum for small farmers in rural India. In: CHI. ACM, pp. 733–742.

Pellegrini, T., Lamel, L., 2006. Investigating automatic decomposition for ASR in less represented languages. In: ICSLP'06, Pittsburgh.

Pellegrini, T., Lamel, L., 2008. Are audio or textual training data more important for ASR in less-represented languages?. In: SLTU'08, Hanoi, Vietnam.

Pellegrini, T., Lamel, L., 2009. Automatic word decompounding for ASR in a morphologically rich language: application to Amharic. IEEE Transactions on Audio, Speech & Language Processing 17 (5), 863–873.

Plahl, C., Schlueter, R., Ney, H., 2011. Cross-lingual portability of Chinese and English neural network features for French and German LVCSR. In: Proc. ASRU, USA.

Rastrow, A., Dredze, M., Khudanpur, S., 2012. Fast syntactic analysis for statistical language modeling via substructure sharing and uptraining. In: Proc. 50th Annual Meeting of Association for Computational Linguistics ACL'2012, Jeju, Korea, pp. 175–183.

Ronzhin, A., Karpov, A., 2007. Russian voice interface. Pattern Recognition and Image Analysis 17 (2), 321–336.

Rotovnik, T., Maucec, M.S., Kacix, Z., 2007. Large vocabulary continuous speech recognition of an inflected language using stems and endings. Speech Communication 49 (6), 437–452.

Roux, J.C., Botha, E.C., du Preez, J.A., 2000. Developing a multilingual telephone based information retrieval system in African languages. In: Proceedings of the Second International Conference on Language Resources and Evaluation, pp. 975–980.

Sak, H., Saraclar, M., Güngör, T., 2010. Morphology-based and sub-word language modeling for Turkish speech recognition. In: ICASSP 2010, pp. 5402–5405.

Sarikaya, R., Afify, M., Gao, Y., 2007. Joint morphological-lexical language modeling (JMLLM) for Arabic. In: Proc. ICASSP'07, vol. 4, pp. 181–184.

Schlippe, T., Ochs, S., Schultz, T., 2010. Wiktionary as a source for automatic pronunciation extraction. In: Interspeech 2010, Makuhari, Japan, 26–30 September 2010.

Schlippe, T., Ochs, S., Schultz, T., 2012a. Grapheme-to-phoneme model generation for indo-European languages. In: ICASSP 2012, Kyoto, Japan, 25–30 March 2012.

Schlippe, T., Ochs, S., Vu, N.T., Schultz, T., 2012b. Automatic error recovery for pronunciation dictionaries. In: Interspeech 2012, Portland, Oregon, 9–13 September 2012.

Schlippe, T., Ochs, S., Schultz, T., 2013. Web-based tools and methods for rapid pronunciation dictionary creation. Speech Communication. http://dx.doi.org/10.1016/j.specom.2013.06.015.

Schultz, T., 2002. GlobalPhone: a multilingual speech and text database developed at Karlsruhe University. In: ICSLP, pp. 345–348.

Schultz, T., 2006. Multilingual speech processing. In: Tanja Schultz, Katrin Kirchhoff (Eds.), Elsevier, Academic Press, ISBN 13: 978-0-12-088501-5, 2006.

Schultz, T., Black, A.W., Badaskar, S., Hornyak, M., Kominek, J., 2007. SPICE: web-based tools for rapid language adaptation in speech processing systems. In: Interspeech 2007, Antwerp, Belgium.

Schultz, T., Vu, N.T., Schlippe, T., 2013. GlobalPhone: a multilingual text & speech database in 20 languages. In: ICASSP 2013, Vancouver, Canada.

Schultz, T., Waibel, A., 1998. Language independent and language adaptive LVCSR. In: Proc. ICSLP, Sydney, pp. 1819–1822.

Schultz, T., Waibel, A., 2001. Language independent and language adaptive acoustic modeling for speech recognition. Speech Communication 35, 31–51.

Seide, F., Li, G., Chen, X., Yu, D., 2011. Feature engineering in context-dependent deep neural networks for conversational speech transcription. In: Proc. ASRU-2011 International Workshop, HI, USA, pp. 24–29.

Siniscalchi, S.M., Reed, J., Svendsen, T., Lee, C.-H., 2013. Universal attribute characterization of spoken languages for automatic spoken language recognition. Computer Speech & Language 27 (1), 209–227.

Solera-Urena, R., Martın-Iglesias, D., Gallardo-Antolın, A., Pelaez-Moreno, C., Dıaz-de-Marıa, F., 2007. Robust ASR using support vector machines. Speech Communication 49 (4), 253–267.

Stahlberg, F., Schlippe, T., Vogel, S., Schultz, T., 2012. Word segmentation through cross-lingual word-to-phoneme alignment. In: Proceedings of The Fourth IEEE Workshop on Spoken Language Technology (SLT 2012), Miami, Florida, 2–5 December 2012.

Stahlberg, F., Schlippe, T., Vogel, S., Schultz, T., 2013. Pronunciation extraction from phoneme sequences through cross-lingual word-to-phoneme alignment. In: Proceedings of the 1st international conference on statistical language and speech processing (SLSP 2013), Tarragona, Spain, 29–31 July 2013.

Stephenson, T.A., Escofet, J., Magimai-Doss, M., Bourlard, H., 2002. Dynamic Bayesian network based speech recognition with pitch and energy as auxiliary variables, Technical Report Idiap-RR-24-2002, p. 10.

Stolcke, A., Grezl, F., Hwang, M.-Y., Lei, X., Morgan, N., Vergyri, D., 2006. Cross-domain and cross-lingual portability of acoustic features estimated by multilayer perceptrons. In: Proc. ICASSP 2006.

Stüker, S., 2008. Integrating Thai grapheme based acoustic models into the ML-mix framework – for language independent and cross-language ASR. In: SLTU'08, Hanoi, Vietnam.

Stüker, S., Schultz, T., Metze, F., Waibel, A., 2003. Multilingual articulatory features, In: ICASSP 2003.

Stuker, S., Schultz, T., Metze, F., Waibel, A., 2003. Multilingual articulatory features. In: Proceedings. ICASSP'03 IEEE International Conference on Acoustics, Speech, and, Signal Processing.

Stüker, S., Besacier, L., Waibel, A., 2009. Human translations guided language discovery for ASR systems. In: InterSpeech-2009, Brighton, UK.

Suenderman, K., Liscombe, J., 2009. Localization of speech recognition in spoken dialog systems: how machine translation can make our lives. In: Interspeech 2009, Brighton, UK, pp. 1475–1478.

Szarvas, M., Furui, S., 2003. Finite-state transducer based modeling of morphosyntax with applications to Hungarian LVCSR. In: Proc. ICASSP, HongKong, China, pp. 368–371.

Tachbelie, M., Abate, S.T., Besacier, L., Rossato, S., 2012. Syllable-based and hybrid acoustic models for Amharic speech recognition. In: SLTU – Workshop on Spoken Language Technologies for Under-Resourced Languages, Cape-Town, South Africa.

Tachbelie, M., Abate, S.T., Besacier, L., 2013. Using different acoustic, lexical and language modeling units for ASR of an under-resourced language – Amharic. Speech Communication. http://dx.doi.org/10.1016/j.specom.2013.01.008.

Tarjan, B., Mihajlik, P., 2010. On morph-based LVCSR improvements. In: Proc. 2nd Int. Workshop on Spoken Languages Technologies for Under-resourced Languages SLTU-2010, Malaysia, pp. 10–16.

Thomas, S., Ganapathy, S., Hermansky, H., 2012a. Multilingual MLP features for low-resource LVCSR systems. In: Proc. ICASSP, Japan.

Thomas, S., Ganapathy, S., Jansen, A., Hermansky, H., 2012b. Data-driven posterior features for low resource speech recognition applications. In: Proc. Interspeech, USA.

Toth, L., Frankel, J., Gosztolya, G., King, S., 2008. Cross-lingual portability of MLP-based tandem features – a case study for English and Hungarian. In: Proc. Interspeech.

Trentin, E., Gori, M., 2001. A survey of hybrid ANN/HMM models for automatic speech recognition. Neurocomputing 37 (1), 91–126.

van Heerden, C., Kleynhans, N., Barnard, E., Davel, M., 2010. Pooling ASR data for closely related languages. In: Proceedings of the Workshop on Spoken Languages Technologies for Under-Resourced Languages (SLTU 2010), Penang, Malaysia, May 2010, pp. 17–23.

van Niekerk, D.R., Barnard, E., 2013. Predicting utterance pitch targets in Yoruba for tone realisation in speech synthesis, Speech Communication. http://dx.doi.org/10.1016/j.specom.2013.01.009.

Vergyri, D., Kirchhoff, K., Duh, K., Stolcke, A., 2004. Morphology-based language modeling for Arabic speech recognition. In: Proc. ICSLP'04, pp. 2245–2248.

Vesely, K., Karafiat, M., Grezl, F., Janda, M., Egorova, E., 2012. The language-independent bottleneck features. In: Proc. SLT, USA.

Vu, N.T., Kraus, F., Schultz, T., 2010. Multilingual A-stabil: a new confidence score for multilingual unsupervised training. In: Proc. SLT, USA.

Vu, N.T., Kraus, F., Schultz, T., 2011. Rapid building of an ASR system for under-resourced languages based on multilingual unsupervised training. In: Proc. Interspeech, Italy.

Vu, N.T., Metze, F., Schultz, T., 2012a. Multilingual bottle-neck feature for under resourced languages. In: Proc. SLTU, South Africa.

Vu, N.T., Breiter, W., Metze, F., Schultz, T., 2012b. An investigation on initialization schemes for multilayer perceptron training using multilingual data and their effect on ASR performance. In: Proc. Interspeech, USA.

Wheatley, B., Kondo, K., Anderson, W., Muthusamy, Y., 1994. An evaluation of cross-language adaptation for rapid HMM development in a new language. In: Proc. ICASSP, Adelaide, pp. 237–240.

Whittaker, E.W.D., 2000. Statistical language modelling for automatic speech recognition of Russian and English. Ph.D. thesis, Cambridge Univ., p. 140.

Whittaker, E.W.D., Woodland, P.C., 2001. Efficient class-based language modelling for very large vocabularies. In: ICASSP-2001, Salt Lake City, USA, pp. 545–548.

Wissing, D., Barnard, E., 2008. Vowel variations in Southern Sotho: an acoustical investigation. Southern African Linguistics and Applied Language Studies 26 (2), 255–265.

Young, S.J., Adda-Decker, M., Aubert, X., Dugast, C., Gauvain, J.L., Kershaw, D.J., Lamel, L., Leeuwen, D.A., Pye, D., Robinson, A.J., Steeneken, H.J.M., Woodland, P.C., 1997. Multilingual large vocabulary speech recognition: the European SQALE project. Computer Speech & Language 11, 73–89.

Young, S., 2008. HMMs and related speech recognition technologies. In: Springer Handbook of Speech Processing. Springer-Verlag, Berlin Heidelberg, pp. 539–557.

Yu, D., Siniscalchi, S.M., Deng, L., Lee, C.-H., 2012. Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition. In: Proc. ICASSP-2012, pp. 4169–4172.