Current trends in multilingual speech processing

HERVÉ BOURLARD[1,2,*], JOHN DINES[1],
MATHEW MAGIMAI-DOSS[1], PHILIP N GARNER[1],
DAVID IMSENG[1,2], PETR MOTLICEK[1], HUI LIANG[1,2],
LAKSHMI SAHEER[1,2] and FABIO VALENTE[1]

[1]Idiap Research Institute, CH-1920 Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
e-mail: herve.bourlard@idiap.ch

Abstract.   In this paper, we describe recent work at Idiap Research Institute in the domain of multilingual speech processing and provide some insights into emerging challenges for the research community. Multilingual speech processing has been a topic of ongoing interest to the research community for many years and the field is now receiving renewed interest owing to two strong driving forces. Firstly, technical advances in speech recognition and synthesis are posing new challenges and opportunities to researchers. For example, discriminative features are seeing wide application by the speech recognition community, but additional issues arise when using such features in a multilingual setting. Another example is the apparent convergence of speech recognition and speech synthesis technologies in the form of statistical parametric methodologies. This convergence enables the investigation of new approaches to unified modelling for automatic speech recognition and text-to-speech synthesis (TTS) as well as cross-lingual speaker adaptation for TTS. The second driving force is the impetus being provided by both government and industry for technologies to help break down domestic and international language barriers, these also being barriers to the expansion of policy and commerce. Speech-to-speech and speech-to-text translation are thus emerging as key technologies at the heart of which lies multilingual speech processing.

Keywords.   Multilingual speech processing; speech synthesis; speech recognition; speech-to-speech translation; language identification.

1. Introduction

Multilingual speech processing (*MLSP*) is a distinct field of research in speech and language technology that combines many of the techniques developed for monolingual systems with new approaches that address specific challenges of the multilingual domain. Research in MLSP should be of particular interest to countries such as India and Switzerland where there are

---

*For correspondence

several officially recognized languages and many more additional languages are commonly spoken. In such multilingual environments, the language barrier can pose significant difficulties in both commerce and government administration and technological advances that could help break down this barrier would be of great cultural and economic value.

In this paper, we discuss current trends in MLSP and how these relate to advances in the general domain of speech and language technology. Examples of current advances and trends in the field of MLSP are provided in the form of case studies of research being conducted at Idiap Research Institute in the framework of international and national research programmes. The first case study presents work in personalized speech-to-speech translation, in which we have made significant advances in developing methods for unsupervised cross-lingual adaptation of hidden Markov model (HMM)-based speech synthesis. In addition, this work has been closely linked with efforts to develop unified models for *automatic speech recognition* (*ASR*) and *text-to-speech synthesis* (TTS), thus highlighting the potential of MLSP to impact broader research topics.

The second case study looks at discriminative methods in ASR and how these have been applied to problems in MLSP. This includes a study of multilingual acoustic models, in particular in the context of *multilayer perceptron* (*MLP*)-based discriminative feature extraction. The major component of this study is the use of hierarchical classification frameworks that have the potential to provide more robust performance while simplifying means to perform cross-language knowledge transfer. The third and final case study looks at the related tasks of language identification and out-of-language detection using ASR.

The paper is organized as follows: in section 2, we present a brief literature review of multilingual speech processing, followed by a study of more recent trends in speech and language technology and how these impact MLSP in section 3. In sections 4, 5 and 6, we present the three case studies on work being conducted at Idiap on personalized speech-to-speech translation, discriminative features in multilingual ASR, and language identification and out-of-language detection, respectively. Finally, in section 7 we provide some insights into future opportunities and challenges in MLSP.

2. Multilingual speech processing

In language, there are normally two forms of communication, namely, spoken form and written form[1]. Depending upon the granularity of representation, both these forms can have different or common representation in terms of basic units. For instance, in spoken form, phonemes/syllables can be considered as the basic unit. Similarly, in the case of (most) written forms, graphemes/characters are the smallest basic units. However, above the smaller units, a word[2] can be seen as a common unit for both spoken and written forms. The word then in turn can be described in terms of the smaller basic units corresponding to the spoken form or written form. For a given language, there can be a consistent relationship between spoken form and written form. However, the relationship may not be same across languages. Addressing the issue of a common basic representation across languages is central to MLSP, although this can raise different challenges in applications such as ASR and TTS.

---

[1]It is important to note that not all languages spoken in the world have both forms. There are languages that have spoken form but no written form.

[2]It should be noted that in case of some languages, the written form of word and the character representation may be considered same; e.g. Mandarin language.

Most commonly, the idea of a unified phonetic lexicon is often used as the binding element in MLSP systems. Such a lexicon is available in the International Phonetic Association (IPA) system (IPA 1999). Most phonetic lexicons are in practice just alternative parameterizations of the IPA symbols. Some IPA symbols are difficult to represent in computers, and are catered to by using the SAMPA[3] and SAMPROSA[4] systems. Although SAMPA was originally designed for just European languages, it has occasionally been extended to others (e.g., Chinese (Chen *et al* 2000)). The worldbet system described in (Hieronymus 1993) is more thorough. There is a one-to-one mapping between, say, ARPABET[5] symbols and IPA symbols; notation aside, the former is just a subset of the latter. A drawback of such a common representation is that it can preclude the possibility of tuning aspects of systems to the particular language. For instance, the C–V syllable structure of Japanese and tonal nature of Chinese would normally be hardwired in a monolingual system, but this may not be feasible in the multilingual case.

In the remainder of this section, we give a brief survey of MLSP in the context of specific applications, concentrating on ASR and TTS.

2.1 *Automatic speech recognition*

In the field of ASR, advances have been largely promoted by the US government via the National Institute of Science and Technology (NIST[6]). Besides running evaluations, one of the the main contributions of this has been to provide databases (Paul & Baker 1992; Kubala *et al* 1994). This has led naturally to an English language focus for ASR research and development. That is not to say that ASR research is English centric; rather, many of the algorithmic advances have been made initially in that language. Such advances have, however, translated easily to different languages, largely owing to the robustness of statistical approaches to the different specificities of languages.

State-of-the-art ASR systems are stochastic in nature and are typically based on HMM. Figure 1 illustrates a typical HMM-based ASR system for English language. From a spoken language perspective, the different components of the HMM-based ASR system are as follows.

(i) Feature extractor that extracts the relevant information from the speech signal yielding a sequence of feature observations/vectors. Feature extraction is typically considered a language independent process (i.e., common feature extraction algorithms are used in most systems regardless of language), although in some cases (such as tonal languages) specific processing should be explored.

(ii) Acoustic model that models the relation between the feature vector and units of spoken form (sound units such as phones).

(iii) Lexicon that integrates lexical constraints on top of spoken unit level representation yielding a unit representation that is typically common to both spoken form and written form such as, word or morpheme.

(iv) Language model that tends to model syntactical/grammatical constraints of the spoken language using the unit representation resulting after integrating lexical constraints.

---

[3]http://www.phon.ucl.ac.uk/home/sampa/
[4]http://www.phon.ucl.ac.uk/home/sampa/samprosa.htm
[5]http://en.wikipedia.org/wiki/Arpabet
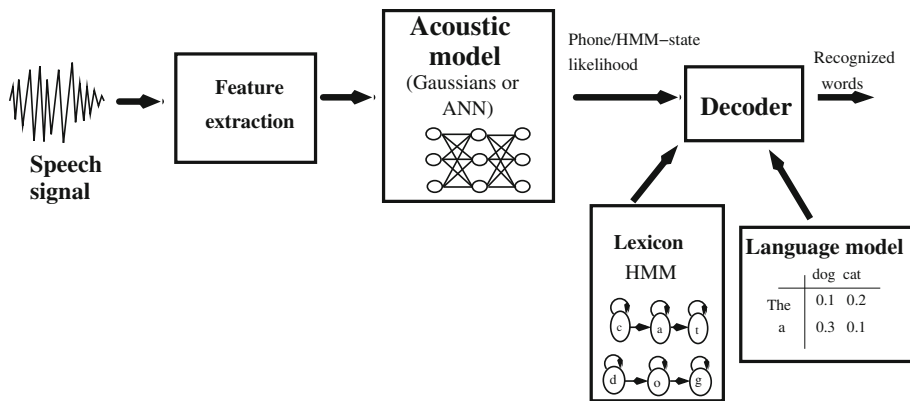[6]http://www.nist.gov/index.html

Figure 1. Illustration of a typical HMM-based ASR system.

The HMM-based ASR system using Gaussians to model the feature observations is referred to as HMM/GMM system (Rabiner 1989). Similarly, hybrid HMM/ANN systems refer to the case where artificial neural networks (ANNs), typically, MLPs, are used to model the feature observations (Bourlard & Morgan 1994) and estimate phone posteriors based on the input features within some temporal context (typically 90 ms).

In ASR, we are generally attempting to address one of two issues in MLSP. The first issue is that of building multilingual models; that is, models that can recognize speech of multiple languages. In building such multilingual systems, a major issue is the availability of resources. At present, there are various resources that enable multilingual speech research. An early example is the OGI multi-language telephone speech corpus (Muthusamy *et al* 1992). More recent ones are GlobalPhone (Schultz & Waibel 1997a), SpeechDat(M)[7] and SPEECON (Siemund *et al* 2000). Complementary to the issue of data resources, the second major area of research in MLSP for ASR concerns cross-language transfer. Specifically, how data resources for a given language(s) can be used to improve ASR in another language, in particular when resources for that language are lacking. Much of the work in cross-language transfer has built upon progress already made in multilingual modelling.

A trivial approach to building multilingual models is to construct separate acoustic and language models for each desired language, i.e., a monolingual system for each language. During recognition, multiple decoders corresponding to the different monolingual systems are run in parallel, and the recognizer output yielding the maximum likelihood is selected. As a by-product, the identity of language is also inferred. Such an approach is simple and feasible. However, in the light of practical issues such as portability to new languages (especially with fewer resources), system memory and computational requirements, it may not be the best approach. Given this, there has been considerable effort devoted to build acoustic models and language models that are shared across languages.

Along the direction of multilingual acoustic modelling, the emphasis has been towards finding a common sound unit representation that is shared across languages (Schultz 2006). In the literature, a popular approach towards this is the creation of a 'universal/global' phone set by first pooling the phone sets of different languages together and then merging them (i) based on heuristics/knowledge such as IPA-based (Köhler 1996) or SAMPA-based (Ackermann

---

[7]http://www.speechdat.org/SpeechDat.html

et al 1996), articulatory features (Dalsgaard & Andersen 1992), (ii) in a data-driven manner by clustering (Köhler 1999) or by measuring phoneme similarity such as using confusion matrix (Andersen *et al* 1993), or (iii) both i.e., knowledge-based merging followed by data-driven clustering (Schultz & Waibel 1998; Köhler 1999). The underlying assumption here is that the articulatory representations of phones are similar across languages, and thus their acoustic realizations can be considered language independent. These studies were mostly done in the framework of modelling context-independent phones. The approach of knowledge-based merging followed by data-driven clustering for context-dependent phone models was further investigated by Schultz & Waibel (2001).

In the literature (Schultz 2006), different studies have been reported for cross-language transfer, where an acoustic model trained on a different language or many different languages (excluding target language) is (i) used directly when no data from the target language is available (Constantinescu & Chollet 1997), (ii) adapted when limited data from target language is available (Wheatley *et al* 1994; Köhler 1998), or (iii) used as seed model and then trained with large amount of data (Osterholtz *et al* 1992; Schultz & Waibel 1997b; Maskey *et al* 2004). When compared to the use of monolingual acoustic models, it has been typically observed that multilingual acoustic models yield a better starting point for such cross-language transfer.

From the above brief literature overview, it is clear that there has been reasonable success in constructing multilingual acoustic models. However, developing multilingual language models is still an open issue (Khudanpur 2006). Standard ASR systems typically use N-gram language models, where word sequences are modelled by a Nth order Markov chain, i.e., the model consists of probability to transit to a word given a word sequence of length N[8]. When estimating N-gram language models for different languages, the language-specific characteristic plays a greater role. For instance, at the word level, the perplexity of the language can vary greatly depending on the degree of morphological inflection. As a result, morphologically rich languages can have many more words, thus leading to morphological-level rather than word-level representations. Also, some morphologically rich languages tend to allow free word ordering, which is clearly not well handled by simple N-gram models. Additionally, character-based languages, such as Chinese and Japanese, do not explicitly identify word boundaries. In such cases, a system for word segmentation is necessary. In spite of all these language-specific issues, there is need for multilingual language models to handle, for example, instances of 'code switching[9]'. Research in the direction of building multilingual language models has mainly focussed on (Weng *et al* 1997; Fügen *et al* 2003; Khudanpur 2006) (i) estimation of a single language model using the data from all languages, or (ii) estimation of a language model for each language separately including the words from other languages and interpolation of these. It has been observed that the latter approach yields better performance than the first approach.

2.2 *Text-to-speech synthesis*

In the domain of TTS, issues in multilingual speech processing have been dominated by two main areas. This is partially due to the limitations of the dominant unit selection paradigm, which directly uses the recordings from voice talents in the generation of synthesized speech. It is clear that multilingual techniques in unit selection are thus bound by the limited availability

---

[8]$N = 0, 1, 2$ refers to unigram, bigram, trigram, respectively.
[9]Code switching refers to the case where the user switches from one language to another language. The switch can happen within an utterance or across utterances.

of multilingual voice talents and more so by the availability of such recordings to the research community, although some research has been conducted to overcome this (Kominek 2009).

The first area concerns the construction of TTS voices in multiple languages, which is dealt with in detail by Sproat (1997). This area of research is largely concerned with the development of methodologies that can be made portable across languages, and include issues in text parsing, intonation and waveform generation. Many of these general issues are mirrored in challenges faced by the multilingual ASR community, as has been discussed in the opening of this section, although many of the related challenges are arguably more complex in TTS due to the extensive use of supra-segmental features and the greater degree of language dependence of such features – there is no 'international prosodic alphabet'. Much progress in this direction has been made owing to efforts of the research community towards the development of freely available corpora and TTS tools (for example, see Festival (Black & Taylor 1997), Festvox (Black & Lenzo 2007) and MBROLA (Dutoit *et al* 1996)), which has enabled the development of synthesis systems in many languages. Although, the systems that have been developed using these resources have remained largely independent of one another, these developments have laid the ground work for future multilingual applications of TTS.

A second area of research that has a more evident emphasis on multilingual capabilities is that of polyglot synthesis (Traber *et al* 1999). In polyglot synthesis, the goal is the synthesis of mixed language utterances, hence, a major challenge is in proper text parsing that enables correct pronunciation and intonation of such speech. Given appropriate text parsing, the main concern of unit selection synthesis is the efficient design of a mixed-language inventory of speech.

A related field of research, voice conversion, has helped overcome some of the limitations of unit selection methods with respect to MLSP. In particular, the application of voice conversion to cross-lignual scenarios has been investigated, especially in the context of speech-to-speech translation (Suendermann *et al* 2006). Drawbacks of voice conversion techniques lie in their limited ability to modify supra-segmental speech characteristics and the requirement of parallel data for learning the conversion, although some progress in this direction has been made.

2.3 *Automatic language recognition*

The objective of automatic language recognition is to recognize the spoken language by automatic analysis of speech data. Automatic language recognition can be classified into two tasks (i) automatic language identification and (ii) automatic language detection. In principle, this classification is similar to speaker identification and speaker verification in speaker recognition research.

The goal of automatic *language identification* (*LID*) is to classify a given input speech utterance as belonging to one out of L languages. Various possible applications of LID can be found in multilingual speech processing, call routing, interactive voice response applications, and front-end processing for speech translation translation. There are a variety of cues, including phonological, morphological, syntactical or prosodic cues, that can be exploited by an LID system (Zissman & Berkling 2001; Navratil 2006). In the literature, different approaches have been proposed to perform LID, such as using only low-level spectral information (Sugiyama 1991), using phoneme recognizers in conjunction with phonotactic constraints (Navratil 2001; Lamel & Gauvain 1993) or using medium- to high-level information (e.g. lexical constraints, language models) through speech recognition (Schultz *et al* 1996). Among these, the most common approach is to use phoneme recognizers along with phonotactic constraints. The phoneme

recognizer can be language-dependent (Lamel & Gauvain 1993) (using a language-specific phoneme set) or language-independent (Berkling & Barnard 1995) (using a multilingual phoneme set). The phonotactic constraints are typically modelled by a phoneme bigram estimated on phonetically labelled data.

Given a segment of speech signal and associated claimed language, the goal of automatic language detection is to verify the claim or, in other words, choose one of the two possible hypotheses. A null hypothesis that the given speech segment belongs to the claimed language or the alternative hypothesis that the given segment does not belong to the claimed language. Usually, this is achieved by training a model corresponding to the null hypothesis using data from the target language, and a separate model corresponding to the alternate hypothesis using data from different languages (Campbell *et al* 2006; Burget *et al* 2009).

3. Recent trends in speech and language processing

We can identify a number of advances in speech and language processing that have significantly impacted MLSP. One of the most important developments has been the rise of statistical machine translation that has resulted in substantial funding and, consequently, research activity being directed towards machine translation and its related multilingual applications (speech-to-speech, speech-to-text translation, etc.). Several notable projects have been pursued in recent years, to mention only a few: *Spoken Language Communication and Translation System for Tactical Use* (TRANSTAC DARPA initiative), *Technology and Corpora for Speech to Speech Translation* (TC-STAR FP6 European project), and the *Global Autonomous Language Exploitation* (GALE DARPA initiative[10]). Research in these projects not only needs to focus on the optimization of individual components, but also on how the components can be integrated together effectively to provide overall improved output. This is not a trivial task. For instance, speech recognition systems are typically optimized to reduce word error rate (WER) (or character/letter error rate, CER/LER, in some languages). The WER measure gives equal importance to the different types of error that can be committed by the ASR system, namely, deletion, insertion and substitution. Suppose, if the ASR system output (i.e., text transcript) is processed by a machine translation system, then deletion error is probably more expensive compared to other two errors as all information is lost. It then follows that the optimal performance with respect to WER may not provide the best possible translated output and vice versa. Indeed, in the GALE project (in the context of translation of Mandarin language speech data to English), it was observed that CER is less correlated with translation error rate when compared to objective functions such as SParseval (parsing-based word string scoring function) (Hillard *et al* 2008).

Similarly, the relatively recent emergence of statistical parametric speech synthesis (Zen *et al* 2009) has resulted in a flurry of new research activities and is contributing to substantial efforts to 'cross-pollinate' ideas between ASR and TTS, as techniques in these two fields have become increasingly interrelated (Ostendorf & Bulyko 2002; Gales & Young 2007). HMM-based TTS has helped accelerate efforts in the development of multilingual TTS by providing a means to easily train synthesis systems for new languages, where adaptive techniques may prove particularly useful (Latorre 2006; Kominek 2009). It is also evident that the dominant source-filter model employed in HMM-based TTS (Koishida *et al* 1994; Kawahara *et al* 2001) may not be

---

[10]See respectively: http://www.darpa.mil/ipto/programs/transtac/transtac.asp; http://www.tc-star.org; http://www.darpa. mil/ipto/programs/gale/gale.asp

ideal for all languages and further work is being carried out to address this problem (Silén *et al* 2009).

There has also been increasing effort in the use of lightly supervised and unsupervised training methods for ASR (Lamel *et al* 2002), which has more recently been applied to TTS (Yamagishi *et al* 2009). Methods that use 'found' data from the internet have also been shown to be useful (Bulyko *et al* 2007; Wan & Hain 2006). Such approaches have the potential to have a great impact on efforts in MLSP for poorer resourced languages or languages with little transcribed material, as has been demonstrated by Lööf *et al* (2009). There have also been efforts in cross-lingual language modelling, where cross-lingual information retrieval and machine translation have been used with (i) sentence-aligned parallel corpus (Kim & Khudanpur 2002), (ii) document-aligned corpus (Kim & Khudanpur 2003), and (iii) latent semantic analysis (Kim & Khudanpur 2004) to improve language modelling for resource-deficient language.

Finally, relatively recent efforts have been introduced to integrate more appropriate training criteria in machine learning algorithms that provide more discriminative models in the case of ASR (Schlüter *et al* 2001) and better quality synthesis for TTS (Wu & Wang 2006). This has not only been restricted to the estimation of model parameters, but has also been applied to feature extraction, such as MLP features (Hermansky *et al* 2000). These methods have been shown to provide significant improvements for monolingual ASR systems, but generalization to the multilingual case has not been extensively investigated. There are different ways for improvement of these methods, such as use of hierarchical methods based on MLP (Pinto 2010) or conditional random fields (Fosler-Lussier & Morris 2008), MLP regularization approach (Li & Bilmes 2006) to handle resource (data) differences across languages through cross-language transfer.

3.1 *Recent and ongoing MLSP research at Idiap Research Institute*

At Idiap, multilingual speech processing is an important research objective given Switzerland's location within Europe, at the intersection of three major linguistic cultures, and Swiss society itself being inherently multilingual. Towards this end, we have been conducting research in MLSP as part of several internationally and nationally funded projects, including for instance:

 (i) **EMIME**[11] (*Effective Multilingual Interaction in Mobile Environments*): This FP7 EU project commenced in March 2008 and is investigating the personalization of speech-to-speech translation systems. The EMIME project aims to achieve its goal through the use of HMM-based ASR and TTS, more specifically, the main research goal is the development of techniques that enable unsupervised cross-lingual speaker adaptation for TTS.
 (ii) **GALE**[12] (*Global Autonomous Language Exploitation*): Idiap was involved in this DARPA-funded project as part of the SRI-lead team. The project primarily involved machine translation and information distillation. We have mostly studied the development of new discriminative MLP-based features and MLP combination methods for the ASR components. Despite cessation of the project, we have continued our research of this topic.
(iii) **MULTI** (*MULTImodal Interaction and MULTImedia Data Mining*) is a Swiss National Science Foundation (SNSF) project carrying out fundamental research in several related fields of multimodal interaction and multimedia data mining. A recently initiated MULTI sub-project is conducting research in MLP-based methods for language identification and multilingual speech recognition with a focus on Swiss languages.

---

[11]www.emime.org. See Wester *et al* (2010) for an overview of the project.
[12]http://www-speech.sri.com/projects/GALE

(iv) **COMTIS**[13] (*Improving the coherence of machine translation output by modelling intersen-tential relations*) is a new machine translation project funded by the SNSF that was started on March 2010. The project is concerned with modelling the coherency between sentences in machine translation output, thereby improving overall translation performance.

In the remainder of the paper, we present three case studies conducted in MSLP at Idiap that have resulted from participation in the above-mentioned research projects.

4. Personalizing speech-to-speech translation

One aspect which we take for granted in spoken communication that is largely missing from current speech-to-speech translation (SST) technology is a means to facilitate the personal nature of spoken dialogue. That is, state-of-the-art approaches lack or are limited in their ability to be personalized in an effective and unobtrusive manner, thereby acting as a barrier to natural communication. The use of a common framework for ASR and TTS provides several interesting research opportunities in the framework of SST, including the development of unified approaches for the modelling of speech for recognition and synthesis that will need to adapt across languages to each user's speaking characteristics. In this section, we present progress that has recently been made by Idiap in the EMIME project, but firstly, we discuss the role of translation in our study.

4.1 *Machine translation as the glue*

Unlike in most other speech-to-speech translation projects, translation plays a less prominent role in EMIME. The key focus of research is the personalization of SST, which essentially requires the development of techniques for unsupervised cross-lingual speaker adaptation for HMM-based TTS. Translation acts as the 'glue' that links the input and output languages for cross-lingual adaptation and also links ASR and TTS for unsupervised adaptation. Thus, we can consider our goal as comprising two main tasks:

- To *bridge the gap between ASR and TTS* by investigating techniques in unsupervised adaptation for TTS.
- To *bridge the gap between languages* such that we can perform cross-lingual adaptation of HMM-based TTS.

We are working with several languages that encompass a range of language families, geo-graphical regions and partner competencies: English, Finnish, Japanese and Mandarin. English always comprises one of the languages in each SST language pair. At Idiap, our research has pri-marily focused on the English–Mandarin language pair. English–Mandarin is arguably the most disparate of the language pairs under consideration. Although, this poses a greater challenge, it may better enable us to ascertain and analyse differences between the different approaches under investigation.

In performing unsupervised cross-lingual speaker adaptation within a speech-to-speech trans-lation framework, we consider two possible approaches: a 'pipeline' framework, in which individual components operate largely independent of one another; and a 'unified' framework in which ASR and TTS modules share common components such as feature extraction and acoustic

---

[13]www.idiap.ch/comtis

models. It should be clear that, while the pipeline framework contains a great deal of redundancy, it allows each component to be separately optimized, whereas the unified framework minimizes redundancy, but possibly at a cost to performance. Figure 2 illustrates these two frameworks.
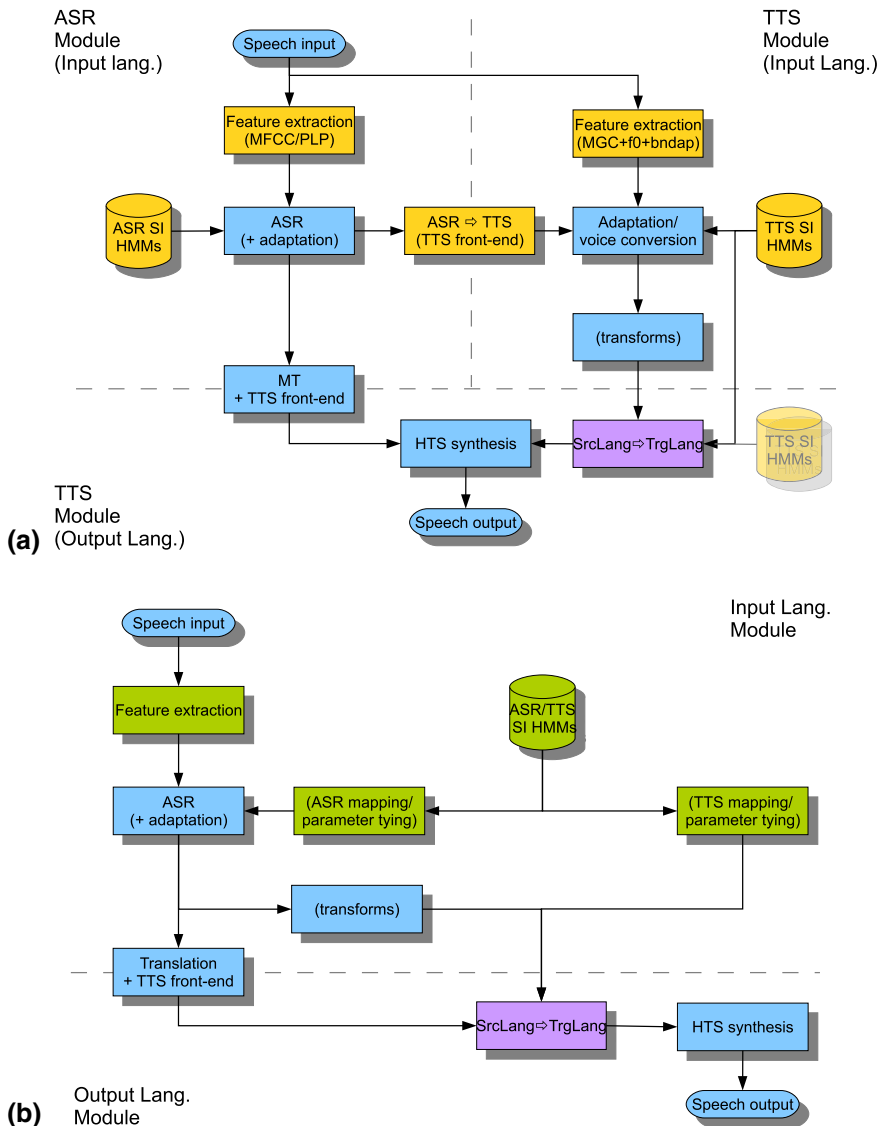

Figure 2. Frameworks for speech-to-speech translation. 'ASR ⇒ TTS' denotes mapping of ASR (triphone context) labels to TTS (full context) labels via a TTS front-end. 'SrcLang ⇒ TrgLang' denotes cross-lingual speaker adaptation (CLSA) from the input language to the output language (see section 4.3 for further details). (**a**) Pipeline approach, whereby ASR and TTS feature extraction and models do not share common components. Separate TTS models may also be employed in input and output languages. (**b**) Unified approach, whereby feature extraction and acoustic models are shared between ASR and TTS and across languages.

4.2 *Bridging the gap between ASR and TTS*

Statistical parametric approaches have emerged in recent years as a dominant paradigm for TTS. Training of such models is very similar to the training of models for ASR – acoustic features are first generated that are used to train acoustic models given corresponding labels. During synthesis, the label sequence is generated from the text to be synthesized. The acoustic model is then used to generate the maximum *a posteriori* probability observation sequence for the given labels, taking into account the explicit relationship between dynamic and static components of the feature vector (Zen *et al* 2009). This can be considered the inverse of the inference procedure carried out in ASR. Other major differences between ASR and TTS include[14] the following

- *Acoustic features*: should provide necessary information to reconstruct the speech signal, normally including pitch and excitation information
- *Labels*: take into account a much broader range of acoustic and prosodic contexts. Such 'full' context labels are normally produced by first parsing text with a TTS front-end; and
- *Models*: normally differ from standard HMM by including explicit duration modelling (hidden semi-Markov model – HSMM (Zen *et al* 2007)) and provide appropriate distributions for modelling of discontinuous features such as pitch (multi-space probability distribution – MSD (Tokuda *et al* 2002)).

To perform unsupervised speaker adaptation of TTS, we use ASR to generate transcriptions that are necessary to adapt the TTS models. In the pipeline approach, ASR transcriptions form the only link between the ASR and TTS modules. These erroneous ASR transcriptions are then fed through a TTS front-end that is used to generate the 'full-context' labels for the adaptation of the TTS models. In the unified approach, acoustic models and features also link the two and adaptation of TTS is carried out implicitly during the adaptation of the ASR models without the need for a TTS front-end.

As noted previously, paradigms for ASR and TTS have undergone a degree of convergence in recent years and now the HMM-based approach is commonly employed for both. As part of our initial studies, we conducted a comparison of feature extraction, acoustic modelling and adaptation for ASR and TTS (Dines *et al* 2010). Although fully unified ASR and TTS models may be sub-optimal, our goal was to quantify the differences between the two that would enable us to better determine where and by how much these approaches differ. We did this by taking ASR and TTS baselines built on a common corpus and then systematically interchanged ASR and TTS components related to lexicon and phone set, feature extraction, model topology and speaker adaptation. Our findings showed that many of the techniques used in ASR and TTS cannot be simply applied to TTS and ASR, respectively without negative consequences. Despite this, we are still interested in unified modelling – not necessarily as a means to explicitly model speech jointly for both ASR and TTS, but rather as a means to transfer knowledge and approaches between the two. We present two examples of this study which relate to feature extraction and acoustic modelling.

Voice tract length normalization for speech synthesis   Vocal tract length normalization (VTLN) is a feature transformation technique that has been used extensively in speech recognition to provide robust and rapid speaker adaptation (Lee & Rose 1998). VTLN operates on the principle that the vocal tract length varies across different speakers and the formant frequency

---

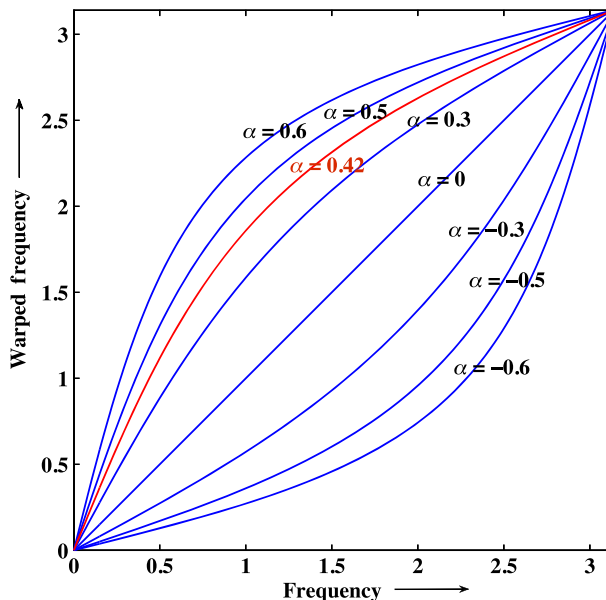[14]Further details can be found in Dines *et al* (2010).

Figure 3. Frequency warping with bilinear transform (0.42 approximates the mel-scale).

positions are inversely proportional to this. Thus, by warping the frequency scale during feature extraction we are able to approximately account for this variability across listeners. An additional advantage of using VTLN in the context of cross-lingual adaptation is that it should be inherently independent of the language being spoken. Building upon earlier studies in ASR, we have been investigating the use of VTLN in statistical parametric speech synthesis.

In the context of speaker adaptive speech synthesis, the most commonly employed feature extraction technique is the mel-generalized cepstrum (MGCEP) (Koishida *et al* 1994). MGCEP analysis uses a bilinear transform to achieve spectral warping that can approximate that of the mel auditory scale (see figure 3). The bilinear transform has also been used to perform VTLN (McDonough 2000) and it follows that a cascade of bilinear transforms simply gives another bilinear transform. Hence, VTLN can be made implicit to MGCEP analysis and can further be formulated as a linear transform in the cepstral domain, which permits optimization via grid-search or expectation-maximization.

We have shown that VTLN for TTS carries with it additional challenges. Firstly, the high-dimensional nature of TTS features results in more severe impact of warping on the feature space, not only making the use of Jacobian normalization imperative (Saheer *et al* 2010b), but also requiring care with the initialization of the model (Saheer *et al* 2010c). Secondly, it is not clear that the ideal criterion for VTLN in TTS is the same as that typically used in ASR. We have noted that the warping factors inferred using the usual objective function do not result in perceptible warping of the voice, while a modified criterion achieves warping that is subjectively closer to the target speaker (but, in contrast, is detrimental to ASR performance) (Saheer *et al* 2010a). This result suggests, once again, the divergent nature of ASR and TTS in terms of direct compatibility, but demonstrates the portability of fundamental techniques between the two.

Decision tree marginalization   In unsupervised adaptation of TTS, we adapt synthesis models from the noisy speech recognition output. If we wish to bypass the TTS front-end, this requires
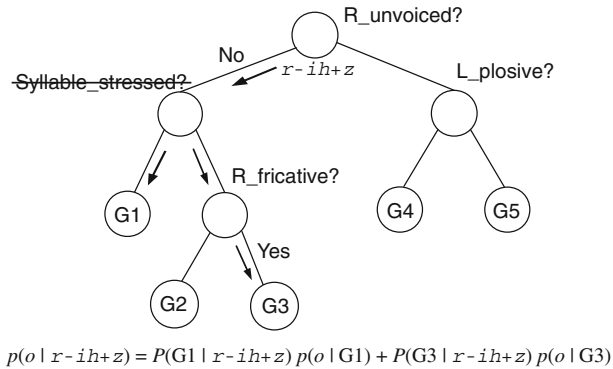
$$p(o \mid r\text{-}ih\text{+}z) = P(G1 \mid r\text{-}ih\text{+}z)\, p(o \mid G1) + P(G3 \mid r\text{-}ih\text{+}z)\, p(o \mid G3)$$

Figure 4. Illustration of decision tree marginalization. In the example, a question 'Syllable_stressed?' is marginalized out of the decision tree. Any distribution of a context that involves traversal of the 'Syllable_stressed?' branch will become the weighted sum of distributions reached by following both children.

a means to adapt the full-context TTS acoustic models from the triphone-context labels generated by the ASR. Several approaches have been proposed to achieve this end, including a method that transfers regression class trees from triphone models to full-context models (King *et al* 2008) and approaches that consider triphone and full-context models using a shared set of parameters (Dines *et al* 2009; Gibson 2009).

We have proposed the *decision tree marginalization* approach that takes a standard set of full-context models trained for TTS and marginalizes out the contexts irrelevant to ASR (e.g., leaving triphone only contexts) (see figure 4). The marginalized models can then be used to estimate adaptation transforms from ASR transcriptions or can be used to directly perform the ASR, although at a cost to performance (Dines *et al* 2009). We have shown that such models can be used in unsupervised adaptation of TTS with minimal impact on synthesis quality compared to supervised adaptation using the full-context models (Liang *et al* 2010).

4.3 *Bridging the gap between languages*

Another aspect of the EMIME project concerns the adaptation of speaker identity across languages for HMM-based TTS. As discussed earlier, multi-linguality for TTS has previously been mostly concerned with polyglot synthesis. Likewise, ASR multilingual modelling has been mostly concerned with building acoustic models that can recognize multiple languages or with cross-language transfer in mind. Thus, cross-lingual speaker adaptation constitutes something of a new task.

We define the input language as the language in which speech input is provided to the SST system and the output language as the language in which spoken output is generated by the SST system. To date, the main approaches that have been considered involve a mapping from models/states in the input language to models/states in the output language (Wu *et al* 2008, 2009). The most successful approaches have involved state-mappings derived from KL-divergence (KLD) between models trained on input and output language data. The mappings can then be used to map either transforms or distributions between languages (see figure 5a).

- In *transform mapping*, intralingual adaptation is carried out in the input language. Generated transforms are then mapped to the output language acoustic models by associating
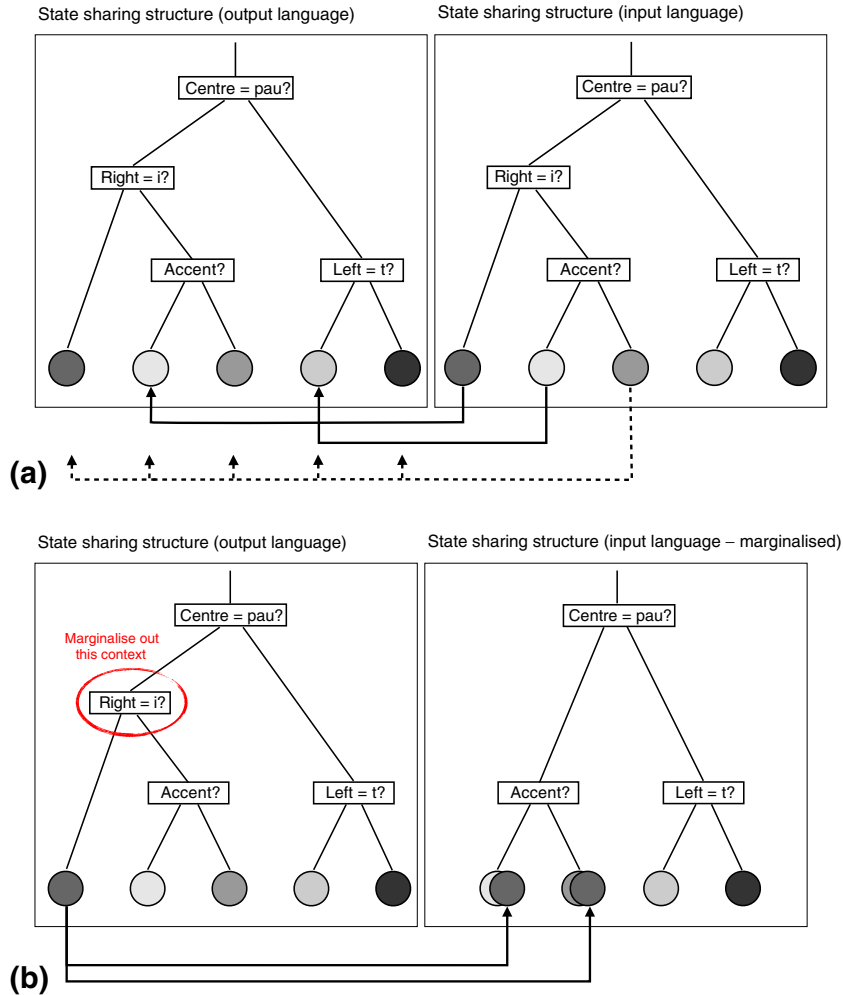
Figure 5. State mapping approaches in which state emission pdfs from acoustic models for the input and output languages are associated with each another. (**a**) State mapping using data mapping (shown) finds a mapping from a state in the output language acoustic model to each state in the input language acoustic model according to minimum KLD. Transform mapping (not shown) uses the same mapping principle, but in the opposite direction, from input language to output language. (**b**) State mapping using decision tree marginalization represents state distributions in the input language as a mixture of state distributions of the output acoustic model.

each HMM state distribution in the output language acoustic model with a state distribution in the input language acoustic model according to the KLD criterion. Transforms are then transferred from the input language states to the output language states according to the defined mapping.

• In *data mapping*, each state in the input language acoustic model is associated with a state from the output language acoustic model according to KLD criterion. Output language states are then substituted for input language states according to this mapping and

intralingual adaptation is performed. The generated transforms may then be directly applied to the output language acoustic model.

We have concentrated on studying these HMM state mapping techniques in both supervised and unsupervised adaptation (using decision tree marginalization) modes of operation (Liang *et al* 2010). We also proposed an alternative *stochastic mapping* approach that uses decision tree marginalization on the output language acoustic models to marginalize out all contexts that are unique to the output language acoustic model, such that the resulting acoustic model can be used directly on the input language (see figure 5b). The results of this first study showed that unsupervised and supervised adaptation using maximum likelihood linear transformations (MLLT) gave similar performance as in the intralingual adaptation scenario. This was a good result as it demonstrated the robustness of adaptation algorithms vis-à-vis unsupervised and cross-lingual scenarios. We also showed that the language of the reference speech plays an important role in people's ability to evaluate speaker similarity, a topic which is now under investigation (Wester *et al* 2010; Wester 2010a,b). Finally, of the different cross-lingual adaptation approaches that were investigated, it was apparent that those using the HMM state emission distributions of the output language acoustic model for transform estimation were preferred over the transform mapping approach.

In light of this final observation, we performed further analysis to ascertain where the state mapping techniques may still be inferior to intralingual approaches specifically studying the influence of language mismatch during adaptation and synthesis (Liang & Dines 2010). We analysed adaptation performance with respect to differing number of transforms and amount of adaptation data, from which it became clear that cross-lingual approaches were not effectively able to benefit from the availability of larger quantities of adaptation data. In particular, the aforementioned data mapping approach was more susceptible to take on characteristics associated with the input language rather than input speaker, resulting in increasingly distorted synthesized speech in the output language as the number of adaptation transforms was increased (see figure 6). This behaviour was attributed to the inherent mismatch between phones in the source and target languages and has made it clear that in order to further improve cross-lingual adaptation, we will need to counter this mismatch during adaptation.
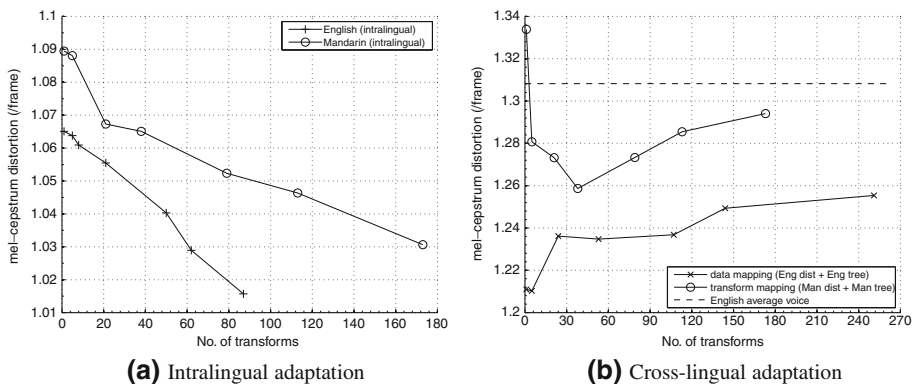

(a) Intralingual adaptation   **(b)** Cross-lingual adaptation

Figure 6. Comparison of objective speech synthesis results for intralingual and cross-lingual adaptation. Intralingual adaptation experiments show consistent reduction in distortion as a greater number of transforms are used, whereas cross-lingual adaptation experiments show the converse. This increase in distortion as the number of transforms increases is attributed to language mismatch between models and transforms.

4.4 *Summary*

We have presented work being conducted at Idiap as part of the EMIME speech-to-speech trans-lation project. In this work we are not only having to deal with the challenges of multilingual modelling for ASR and TTS, but are addressing more fundamental issues concerning unified modelling. Our results so far indicate that direct attempts at unified modelling do not necessar-ily present themselves as a realistic alternative to separate ASR and TTS modelling, but several promising directions have emerged when unified modelling has been considered as a means to port knowledge and approaches between ASR and TTS. Furthermore, our efforts towards unsu-pervised cross-lingual adaptation have shown that good results can be achieved by using largely conventional approaches. There have been shortcomings in the current approaches but efforts are underway to overcome these issues.

5. MLP features and multilingual ASR

There has been sustained interest in using MLP-based discriminative features for automatic speech recognition for several reasons.

(i) Discriminative nature of the features
(ii) Ability of MLP to handle different types of features and long temporal context at the input.
(iii) Robustness towards speaker (Zhu *et al* 2004) and environmental variation (Ikbal 2004)
(iv) Ability to combine multiple feature streams at MLP output level using probabilistic methods (Misra *et al* 2003; Ikbal *et al* 2004; Valente 2009)
(v) Performance improvements obtained are scalable across different training criteria (Zheng *et al* 2007), as well as with amount of data (Fousek *et al* 2008; Valente *et al* 2010).

Figure 7 depicts a typical MLP feature-based ASR system.

The main components of MLP feature extraction are (i) an MLP trained to classify phonemes/phones and (ii) Karhunen Loeve transformation (KLT) matrix (estimated on a data other than test data) to decorrelate the feature vectors. Optionally, during KLT dimensionality,
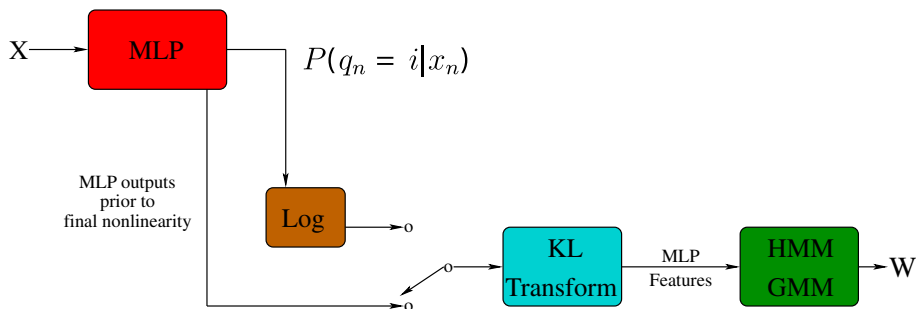

Figure 7. Block diagram of ASR system based on stand-alone MLP features. $X = \{x_1, \cdots x_n, \cdots x_N\}$ represents acoustic feature sequence of length N. W represents the recognized word sequence. $P(q_n = i|x_n)$ represents the a posteriori probability of phone class $i \in \{1, \cdots I\}$ estimated by MLP at time frame n given acoustic feature vector x_n, where I is the number of phone classes or output units of MLP. KL transform refers to Karhunen Loeve transform, which can be either applied to the log of the MLP output vectors or (more or less equivalently) to the MLP output values before the nonlinear (sigmoid/softmax) function.

reduction could be done. This helps in controlling the dimensionality of the feature space, especially when the MLP features are concatenated with the standard spectral features.

Traditionally, a HMM/GMM system models acoustic features that are extracted from short-term spectrum (usually 20–30 ms) of speech signal. The extraction of these acoustic features is assumed to be language-independent (Schultz 2006). When compared to GMM-based modelling, MLPs have the capability to model higher dimensional feature vector (e.g., standard spectral features with temporal context). Furthermore, MLPs avoid the need to make assumption about parametric distribution of input features. As a result, the use of MLP features has led to exploration of spectro-temporal acoustic features (Morgan *et al* 2005; Hermansky & Fousek 2005), i.e., features that span across and characterize both time and frequency. The time span or temporal context in this case can vary from about 250 ms to 1 s, which is about the duration of a syllable. In literature, in the context of MLP feature, it has been found that a combination of spectro-temporal feature processing and conventional spectral processing can lead to a better ASR system. However, as the acoustic–phonetic relationship can differ across languages, one may ask if such spectro-temporal speech processing techniques can also be considered language-independent. Along this line, we present ASR studies using hierarchical MRASTA features for two different languages in section 5.1.

An interesting aspect of MLP feature extraction is that the MLP can be trained on the data that is different from the domain of the task (Sivadas & Hermansky 2004), while still yielding good generalization performance. In the context of multilingual processing, this aspect can be effectively used by training the MLP on a language which has more resources (in terms of data), and using the MLP for languages that have fewer or no resources. This is similar to cross-lingual transfer using acoustic models (discussed earlier in section 2.1), except that cross-lingual transfer here is achieved at feature extraction level. Section 5.2 presents a study on cross-lingual transfer using MLP features.

Similar to multilingual acoustic modelling, the MLP can be trained with data from different languages to classify a 'universal/global' phone set. By sharing data from different languages, such an approach not only helps in handling data issues related to multilingual ASR, but could also help in extracting MLP features that yield a better multilingual ASR system. We present one such recent study in section 5.3.

5.1 *MLP features*

In this section, we present a study using multi-resolution RASTA (MRASTA) feature and hiearchical MRASTA feature (hier-MRASTA). These features were first investigated for English language ASR system, and then extended to Mandarin language ASR system. These studies were originally conducted as part of the DARPA GALE project. More description and details can be found in (Valente & Hermansky 2008; Valente *et al* 2009).

In MRASTA feature extraction (Hermansky & Fousek 2005), first critical band auditory spectrum is extracted through short-term analysis of the speech signal. The number of critical bands depends upon the bandwidth of the speech signal. On Bark scale, there are 15 and 19 critical bands for speech signal of bandwidth 4 kHz and 8 kHz, respectively. A 600 ms long temporal trajectory of each critical band auditory spectrum is then filtered by a bank of filters, also referred to as MRASTA filters. The MRASTA filters are first- and second-order derivatives of Gaussian filters with different variance/time width. In essence, MRASTA filters are multi-resolution band-pass filters on modulation frequency. Finally, approximate derivatives across three consecutive critical bands are computed. An MLP is then trained to classify phones/phonemes using these features as input.

Table 1. Comparing stand-alone MLP feature hier-MRASTA, MLP feature MRASTA, and standard cepstral features across two different languages, namely, English and Mandarin. The performance of English ASR system is expressed in terms of WER on NIST RT05 evaluation data, whereas, the performance of Mandarin ASR system is expressed in terms of CER on GALE Mandarin 2006 evaluation data.

English			Mandarin		
PLP	MRASTA	hier-MRASTA	MFCC	MRASTA	hier-MRASTA
42.4%	45.8%	40.0%	29.9%	32.4%	27.8%

In hierarchical MRASTA feature extraction (Valente & Hermansky 2008) instead of training a single MLP, the filter banks are split into two parts. The first part extracting high modulation frequencies (above 10 Hz), and the second part extracting low modulation frequencies (below 10 Hz). The higher and lower modulation frequencies are then processed in a sequential fashion using a hierarchy of MLPs. More specifically, the first stage MLP processes high modulation frequencies, and the second stage MLP jointly processes low modulation frequencies along with MLP features extracted using first stage MLP. For details, the reader is referred to (Valente *et al* 2009).

The MRASTA and hier-MRASTA features were first studied for English language. The training data consisted of 112 hours of meeting data from different sites (Valente & Hermansky 2008). Thirty-nine-dimensional PLP cepstral features consisting of 13 static coefficients, their approximate first-order and second-order derivatives was used as baseline feature. The MLP of MRASTA feature extractor and MLPs of hier-MRASTA feature extractor were trained to classify 45 context-independent phonemes. The HMM/GMM system was trained using HTK with maximum likelihood criterion. The ASR systems were tested using NIST RT05 evaluation data. The resulting WER for the systems using PLP feature, MRASTA feature and hier-MRASTA feature are shown in table 1.

For Mandarin language ASR studies, we used 100 hours training data setup consisting of broadcast news and broadcast conversation data. The spectral feature baseline system was trained with 39-dimensional MFCC[15] feature vector consisting of 13 static coefficients (extracted after vocal tract length normalization), their approximate first-order and second-order time derivatives. The MRASTA MLP and hier-MRASTA MLPs were trained to classify 71 context-independent phonemes (with tone). During MLP feature extraction (i.e., during KLT), the dimension of MLP feature was reduced to 35. The studies were conducted using SRI/UW/ICSI Mandarin system with maximum likelihood training criteria. The ASR systems were tested using GALE Mandarin 2006 evaluation data. Table 1 presents the the performance of the three systems in terms of CER.

It can be observed that MRASTA features consistently yield the lowest recognition performance across both the languages, while hier-MRASTA features consistently yield a better system compared to the standard spectral feature PLP (in the case of English) and MFCC (in the case of Mandarin). These results tend to show that the trends of MLP features MRASTA and hier-MRASTA, which span longer temporal context, can generalize across languages. The observation about generalization to other languages is further supported by studies reported in

---

[15]In Valente *et al* (2009), the baseline system was trained with 42-dimensional feature vector consisting of 39-dimensional MFCC features and log pitch frequency and its approximate first- and second-time derivatives. We dropped the log pitch frequency features as English study did not use these features.

the literature for languages such as English, Mandarin and Arabic; where it has been observed that MLP features (including MRASTA, hier-MRASTA, and similar processing techniques) are complementary to standard short-term spectral-based features (Morgan *et al* 2005; Fousek *et al* 2008; Valente *et al* 2009). In other words, a system trained with standard cepstral features concatenated with MLP features consistently yield a better performance than a system trained with only standard cepstral feature.

It has to be noted that while multilingual ASR systems aim to use the feature set that is most language-independent, this feature set may not be the best set for an individual language, i.e., there may be features that are more specific to a language or more specifically applicable to a language. For instance, Mandarin is a tonal language; in this case, it has been observed that using pitch frequency features in addition to cepstral feature usually leads to better performance (Valente *et al* 2009).

5.2 *Cross-lingual MLP features*

MLP features are extracted by projecting spectral features along linguistic dimensions, while the projection is 'trained/learned' from data. Given this, they can be applied to transfer knowledge across domains or languages, especially for target domains or languages where less amount of data or no data is available. In this section, we present a cross-lingual feature study, where the MLP is trained on one language and used for feature extraction in another language. This study was originally conducted as an extension of JHU WS06[16], and as well as part of DARPA GALE project.

The study was conducted on Mandarin language. The training data consisted of 97 hours of broadcast news, specifically LDC Mandarin Hub4 and TDT4. The GALE 2004 Mandarin rich transcription development and evaluation sets were used for tuning and evaluating the system, respectively. A monolingual MLP was trained on the Mandarin data to classify 65 Mandarin phones (with tone). A cross-lingual MLP was trained on 2000 hours of conversational telephone speech data of English language to classify 46 context-independent phones. For both MLPs, the spectral features used were 39-dimensional PLP cepstral coefficients with nine-frame temporal context. For more details about the experiment, the reader is referred to Cetin *et al* (2007).

Table 2 shows the performance of three systems, (i) using only MFCC features, (ii) using MFCC features appended with tandem features extracted from the Mandarin MLP (referred to as monolingual tandem), and (iii) using MFCC feature appended with tandem feature extracted from the English MLP (referred to as cross-lingual tandem). It can be observed that both monolingual and cross-lingual MLP features in concatenation with MFCC feature lead to improvement in the ASR performance. It can also be noted that the improvement using crosslingual MLP features is not as significant as monolingual MLP features. This could be because Mandarin and English are very different languages, i.e., English phonetic space may not represent well the Mandarin phonetic space. Further, it should be noted that the Mandarin MLP was trained with speech signal of bandwidth 8 kHz whereas the English MLP was trained with speech signal of bandwidth 4 kHz. Nevertheless, these results suggest that through MLP features, speech data of other languages could be effectively utilized to improve ASR performance of another language.

In the literature, similar cross-lingual studies have been reported. In Stolcke *et al* (2006), it was shown that using MLP-based features extracted from English-trained MLP could improve

---

[16]http://www.clsp.jhu.edu/ws2006/groups/afsr

Table 2. Performance of Mandarin ASR systems investigated in the cross-lingual feature study. Monolingual tandem refers to MLP feature extracted using MLP trained on Mandarin data. Cross-lingual tandem refers to MLP feature extracted using MLP trained on English data. The performance is measured in terms of CER.

Feature	CER
MFCC	21.5%
MFCC+Monolingual tandem	19.5%
MFCC+Cross-lingual tandem	21.2%

Mandarin and Arabic ASR performance over the spectral feature baseline system. In a more recent study (Toth *et al* 2008), cross-lingual portability of MLP features from English language to Hungarian language was investigated by using English-trained phone and articulatory feature MLPs. In addition, a cross-lingual MLP adaptation approach was investigated where the input-to-hidden weights and hidden biases of the MLP corresponding to Hungarian language were initialized by English-trained MLP weights, while the hidden-to-output weights and output biases were initialized randomly. The MLP was then trained on Hungarian data to classify Hungarian context-independent phones. It has to be noted that in essence this cross-lingual MLP adaptation approach is similar to the regularization approach proposed for MLP (Li & Bilmes 2006), where the input-to-hidden mapping is kept intact and the hidden-to-output mapping is relearned. The ASR studies showed that cross-lingual adaptation approach often yields the best system even when compared to the case where the MLP feature is extracted using monolingual MLP (i.e., trained only on Hungarian data).

Although, the studies on cross-lingual MLP features are limited, it has been typically found (including the study presented here) that using MLPs trained on a different language 'directly' may not yield a system better than MLP trained on target language data (if available). In other words, to make better use of the MLPs trained on a different language, cross-lingual adaptation or some kind of training on the target language may be necessary. The cross-lingual adaptation approach discussed earlier (Toth *et al* 2008) is one way this could be achieved. Another way would be to use the recently proposed hierarchical MLP-based phone posterior estimation approach, where two stages (hierarchy) of MLPs are trained to classify context-independent phones (Pinto *et al* 2008, 2011; Pinto 2010). In the first stage, the input feature to the MLPs is a standard spectral-based feature. The input to the second stage MLP is phone posterior probabilities estimated by the first stage MLP with temporal context of around 150–230 ms. This approach has been shown to yield better phoneme recognition performance as well as ASR performance than the single MLP-based approach. In the context of cross-lingual adaptation, the first MLP can be trained on a resource rich language(s) and the second MLP can be trained on the target language with the available data.

5.3 *Multilingual MLP features*

Cross-lingual MLP feature extraction considers training the MLP on a secondary language that has more resources. In the context of multilingual speech recognition, it is possible to consider an MLP trained to classify a universal/global phone set (instead of phone set belonging to a particular language) using data from different languages. Similar to the case of multilingual acoustic modelling, it can be expected that such an approach can help in sharing data from

Table 3. Number of available utterances (utt.), and total duration in hours (h), for each of the five involved languages. English (EN), Spanish (ES), Italian (IT), Swiss French (SF) and Swiss German (SZ).

Lang.	training		dev		test		total	
	utt.	h	utt.	h	utt.	h	utt.	h
EN	3512	1.2	390	0.1	1305	0.4	5207	1.7
ES	3932	1.4	438	0.2	1447	0.5	5817	2.0
IT	3632	1.5	416	0.2	1368	0.6	5416	2.3
SF	3809	1.4	430	0.2	1429	0.5	5668	2.1
SZ	3862	1.3	432	0.1	1426	0.5	5720	1.9
total	18747	6.8	2106	0.8	6975	2.5	27828	10.0

different languages, and can also yield a compact and better multilingual ASR system. In this case, we refer to the MLP as a *multilingual MLP*, and the resulting features as *multilingual MLP features*.

In a preliminary study, we investigated the multilingual MLP features on five European languages, namely, English, Italian, Spanish, Swiss French, and Swiss German from the Speech-Dat(II) corpus (Höge *et al* 1999). The data corresponding to the isolated/application words was used for this study. Table 3 shows the data distribution for different languages. We used the dictionary (based on the SAMPA phone set) provided along with the database. Table 4 shows the number of context-independent phones, and the number of application words (size of lexicon) for each language.

We trained a monolingual MLP corresponding to each language classifying their respective context-independent phones. We adopted the knowledge-driven approach for universal phone set creation, i.e., the phone sets of all the five languages were pooled together and then merged based on their SAMPA symbols. This resulted in a universal phone set with 92 phones (including silence). A multilingual MLP with 39-dimensional PLP cepstral features and nine frames of temporal context as input was then trained to classify this universal phone set. We investigated the following systems for MLP features.

(i) *Mono-tandem*: For each language, a separate acoustic model is built using PLP cepstral features concatenated with MLP feature extracted from their respective monolingual MLP.

(ii) *Multi-tandem*: The multilingual MLP is used here as feature extractor. For each language, the KLT statistics was estimated using only the data specific to the language, and then

Table 4. Information about the languages used in the experiments. The codes are assigned by SpeechDat. The number of phonemes is given based on the reduced lexicon of the application words (not all the phonemes of a language are used).

Language	Code	No. of phonemes (P)	No. of words
British English	EN	33	31
Spanish	ES	29	30
Italian	IT	35	29
Swiss French	SF	36	47
Swiss German	SZ	46	45

Table 5. Word accuracy of the different systems investigated is presented for the two tasks, mono and mixed. Column 4 (rel. loss) presents the relative difference in the performance of the system computed across mono task and mixed task. Mono refers to monolingual speech recognition task. Mixed refers to multilingual/mixed language speech recognition task.

System	Mono	Mixed	Rel. loss
Mono-tandem	98.7%	77.2%	22%
Multi-tandem	98.8%	82.9%	16%
Shared-tandem	97.2%	95.3%	2%

while applying KLT the dimensionality was reduced to match the output dimension of the corresponding monolingual MLP. This dimensionality was reduced to make the system comparable to mono-tandem in terms of complexity. A separate acoustic model was then built for each language separately using the PLP cepstral features concatenated with the multilingual MLP features.

(iii) *Shared-tandem*: Similarly to the multi-tandem system, we used the multilingual MLP for MLP feature extraction. However, in this system, data from all the languages was used for KLT statistics estimation, thus yielding a multlingual MLP feature different from multi-tandem system. In addition, we used the data from all the languages to train a common acoustic model that is shared across languages. In other words, both MLP feature extraction and acoustic modelling are language-independent. It should be noted that here again the feature observation for acoustic model consists of the PLP cepstral features concatenated with multilingual MLP features.

We evaluated the above systems on two different tasks.

(i) *Mono-lingual task*: In this case, it is assumed that the language identity is known *a priori*, and the ASR system corresponding to the language is used for decoding the test utterance. In other words, this task corresponds to monolingual speech recognition.

(ii) *Mixed language task*: In this case, it is assumed that the language identity is not known *a priori*. While decoding the test utterance, all the five ASR systems are run in parallel and the output hypothesis yielding maximum likelihood is selected as the recognized output[17]. In other words, this task corresponds to multilingual speech recognition.

In the case of the mixed language task, it can be observed that mono-tandem and multi-tandem systems have different complexities, i.e., the dimensionality of the feature vectors is different across languages. To handle this, a recognizer dependent bias was subtracted from the respective log likelihood scores (similar to Zissman (1996)) before making a decision about the word hypothesis. The recognizer dependent bias was estimated on the development set.

In table 5, we show the performances for the different systems and tasks. The performance of each system is expressed as the average word accuracy computed across the five languages.

The results show that multilingual MLP features yield the best performance in terms of relative loss in performance between mono and mixed tasks. Although, the shared-tandem system yields

---

[17]As the ASR system in this study is built to recognize isolated words, in case of shared-tandem system it amounts to running a single system.

slightly inferior performance compared to other systems on the mono task, it yields significantly better performance on the mixed task. A similar trend has been previously reported in the context of language independent acoustic modelling. In summary, the superiority of the shared-tandem system on the mixed task can be attributed to the combination of two factors: (i) sharing of data across languages which results in better acoustic model, and (ii) use of multilingual MLP features.

5.4 *Summary*

In this section, we presented three studies on multilingual ASR using MLP features. In the first study, we investigated the MRASTA and hierarchical MRASTA features across two different languages. We found the trends to be similar across languages. The second study presented the use of MLP features for cross-lingual transfer (without any adaptation or retraining), where we found that it is possible to obtain improvements using cross-lingual MLP features. Finally, we presented a preliminary study on the use of multilingual MLP features. Our studies indicate that shared multilingual MLP feature extraction yields better performance when compared to language-specific multilingual MLP feature extraction or monolingual MLP feature extraction.

6. Language identification/detection

In section 2.3, we briefly described that language identification systems use different levels of abstraction related to spoken language processing, including the use of phonotactic constraints, lexical constraints, or both lexical and language constraints through ASR system. Section 6.1 presents a preliminary study on hierarchical MLP-based LID system. This system tries to capture implicitly phonotactic constraints and acoustic confusions present at the output of a multilingual MLP to achieve language identification.

Another way of framing the language detection (LD) problem is in terms of out-of-vocabulary (OOV) detection. Monolingual automatic speech recognition systems assume that the test utterances contain only words from the target language. However, it is possible that segments of test utterances can contain words from foreign language(s), especially in natural conversations. In section 6.2, we present an approach to detect such out-of-language segments using confidence measures.

6.1 *Hierarchical MLP-based language identification*

In section 5.2, we briefly described the recently proposed hierarchical MLP-based phoneme posterior estimation approach and discussed about the potential of applying it for cross-lingual MLP feature extraction. In Pinto *et al* (2011) and Pinto (2010), we have studied the role of the second MLP layer in such hierarchical arrangements using Volterra series and have found that it is predominantly responsible for learning phonetic-temporal patterns present in the posterior features. The learned phonetic-temporal patterns consist of acoustic confusions among phonemes and phonotactic constraints of the language.

In the context of LID, such phonetic-temporal patterns could possibly be exploited by first training an MLP to classify the previously described universal phoneme set (multilingual speech units), and then modelling a larger temporal context of the resulting posterior features by a second MLP to classify languages. It can be expected that information related to phonotactic

Table 6. Comparison of different LID systems. The System Hier performance was obtained with a temporal context of 290 ms at the input of the second stage MLP.

System	Errors	LID %
PC	1236	82.3
SR	360	94.8
Hier	248	96.4

constraints and acoustic confusion among phonemes (present in the posterior features spanning a long temporal context) is language-specific.

We performed a preliminary study on hierarchical MLP-based LID system using the five European language setup and the multilingual MLP described earlier in section 5.3. The second stage MLP of the hierarchical MLP-based LID system was trained with the posterior features (universal phone posterior probabilities) estimated at the output of the multilingual MLP. The temporal context at the input of the second MLP was varied from 130–310 ms. During testing, the decision about the language identity was made by choosing the language that scores the highest log posterior probability over the whole test utterance. We refer to this system as *Hier*.

We compared the hierarchical MLP-based LID system against two different reference systems. In both these systems, the phone posterior probabilities estimated at the output of the multilingual MLP are used as a local score. In the first system, the language-specific phoneme recognizers are run in parallel with their respective bigram phonotactic language model. The LID is achieved by selecting the decoder output that yields the maximum likelihood. This system is referred to as *PC*. In the second system, the language identity is inferred through speech recognition. The second system is similar to the shared-tandem system described earlier in section 5.3 where the acoustic model is shared across languages. However, the recognition is done by using hybrid HMM/MLP-based isolated word recognition system. This system is referred to as *SR*.

Table 6 shows the performance (measured in terms of percentage accuracy) of the different LID systems investigated. In the case of System Hier, the performance is reported for the temporal context of 290 ms. For further details on the effect of temporal context, the reader is referred to Imseng *et al* (2010).

It can be seen that the hierarchical LID system, i.e., Hier system yields the best performance. When comparing the performance of SR and PC systems, the trend is similar to what has been previously reported in the literature. More specifically, higher LID performance has been typically reported (although for fewer number of language classes) using the large vocabulary continuous speech recognition (LVCSR) system when compared to acoustic-phonotactic based systems. Overall, the study shows that there is good potential in exploiting the hierarchial MLP-based approach for language identification.

6.2 *Out-of-language detection*

In multilingual speech processing, the speech data can contain words from different/multiple languages. Earlier in section 2.1, we mentioned that in the context of multilingual ASR this can be possibly handled by the use of multilingual language models. In contrast, there are cases where the goal is to perform monolingual speech recognition but the speech data may contain words (or a sequence of words) from foreign language. For instance, it has been observed that

in spontaneous meeting recordings, the interchangeable use of different languages in short time periods by the same speaker can often be registered (Motlicek 2009). The existence of such segments from foreign language can have an adverse effect on the performance of the ASR system. The adverse effect may be limited by the detection of out-of-language (OOL) segments.

We have proposed a new approach to detect OOL segments through the use of word- and phone-based confidence measures (Motlicek 2009; Motlicek & Valente 2010). In principle, OOL detection can be compared to LD task. However, unlike the LD task, in our OOL detection approach no data from other languages is used. Instead, given a test utterance/segment the OOL detection is achieved by:

(i) Running LVCSR system of the target language to obtain phone lattices or word lattices.
(ii) Treating the lattices as the model and estimating frame level phone posterior probabilities or word posterior probabilities by using standard forward–backward algorithm.
(iii) Estimating a posterior-based confidence measure (CM) from the posterior probability estimate of either phones or words. This is followed by incorporation of temporal context via median filtering of the CM.
(iv) Finally, using the posterior-based CM as threshold on the individual speech segments of the one-best hypothesis obtained from the LVCSR system.

In our work, we have investigated different types of confidence measures and their combination using maximum entropy classifier.

We evaluated the OOL detection technique on Klewel meeting recordings[18]. The evaluation data consists of 3 h of recordings each from three languages, namely, English, French and Italian, i.e., in total 9 h of recordings. English recordings represent in-language speech segments, while French and Italian recordings represent out-of-language segments. This data was processed by an English LVCSR system to obtain word and phone recognition lattices. Experiments on the detection of OOL segments (caused by the French and Italian recordings) yield performances of about 11% EER. Subsequent incorporation of temporal context significantly increased the achieved performance. Median filter with a length of 3 s yield relative improvement of about 62% with respect to the system without application of temporal context.

6.3 *Summary*

In this section, we first presented a LID system based on hierarchical MLP-based approach. Through preliminary studies, we demonstrated that this system could yield better performance than standard approaches, such as modelling phonotactic constraints. We next presented an out-of-language detection approach using confidence measures similar to out-of-vocabulary word detection, and showed its application on real word data.

7. **Future opportunities and challenges**

We have presented an overview of multilingual speech processing – past progress and current trends – from the perspective of our own research activities at Idiap Research Institute. We

---

[18]http://www.klewel.com

have shown that a prime mover behind current trends has been the rise of statistical machine translation, which has had a ripple on effect on the general field of MLSP. It also should be apparent that future trends will still closely follow on from developments made in mainstream speech and language technologies, but the distinct challenges of MLSP will also give rise to novel solutions.

Similar to the influence of statistical machine translation on developments in MLSP in recent years, we anticipate future activity will be strongly driven by web-based services, especially those for mobile devices. We note that these services are becoming widely available and, combined with affordable broadband wireless access, such services will provide an opportunity to make available a broader range of capabilities to mobile devices, especially those based on computationally demanding tasks in MLSP. Thus, we are already seeing services being provided by major market players in the domain of speech processing and we can expect that will also expand to a greater number of applications in MLSP and consequently an increase in research and development activity in both academic and industry alike.

In our own work, two primary research directions that have emerged are cross-lingual speaker adaptation for HMM-based TTS and hierarchical architectures for discriminative MLSP. Work on cross-lingual speaker adaptation for HMM-based TTS has only just started to scratch the surface and is apparent that the rise of statistical parametric TTS will likely lead to many more novel challenges in MLSP for TTS. For example, we have already spoken of joint optimization of combined systems for speech recognition and machine translation. Conceivably, similar principles could be applied to combined machine translation and speech synthesis to produce more intelligible translated output. The adaptive HMM-based framework also poses an attractive solution for research in polyglot synthesis, without the need for developing extensive data resources for multilingual speakers. Addressing the tasks of cross-corpus normalization and cross-language contextual modelling will likely be challenges to overcome if we are to be successful in this.

In the domain of multilingual ASR, methods of cross-language knowledge transfer still have considerable potential. With increasing use of lightly supervised techniques and data mining, there is also increasing need to be able to effectively bootstrap models from other languages. Unfortunately, models trained using discriminative criteria are particularly susceptible to transcription errors, possibly making them unsuitable for application in acoustic model bootstrapping. By combining hierarchical approaches with discriminative techniques, we may obtain an effective technique for acoustic model bootstrapping. Furthermore, in the context of MLP-based features, there is also need to investigate extensively the use of other language-independent representation of phonetic information, such as articulatory features, and modelling of subword unit representations such as graphemes, especially Roman alphabets which is shared across many different languages.

References

Ackermann U, Angelini B, Brugnara F, Federico M, Giuliani D, Gretter R, Lazzari G, Niemann H 1996 Speedata: Multilingual spoken data-entry. In *Proc. Conf. Int. Spoken Lang. Process. (ICSLP)*, pp. 2211–2214

Andersen O, Dalsgaard P, Barry W 1993 Data-driven identification of poly- and mono-phonemes for four European languages. In *Proc. European Conf. Speech Commun. Technol. (EUROSPEECH)*, pp. 759–762

Berkling K, Barnard E 1995 Theoretical error prediction for a language identification system using optimal phoneme clustering. In *Proc. European Conf. Speech Commun. Technol. (EUROSPEECH)*, pp. 351–354

Black A W, Lenzo K A 2007 Building synthetic voices. http://festvox.org/bsv/

Black A W, Taylor P A 1997 The Festival Speech Synthesis System: System documentation. Technical Report HCRC/TR-83, Human Communciation Research Centre, University of Edinburgh, Scotland, UK. Available at http://www.cstr.ed.ac.uk/projects/festival.html

Bourlard H, Morgan N 1994 *Connectionist speech recognition – A hybrid approach*, (Kluwer Academic Publishers, USA)

Bulyko I, Ostendorf M, Siu M, Ng T, Stolcke A, Çetin O 2007 Web resources for language modeling in conversational speech recognition. *ACM Trans. Speech Lang. Process.* 5(1): 1–25

Burget L, Fapso M, Valiantsina H, Glembek O, Karafiat M, Kockmann M, Matejka P, Schwarz P, Cernoky J 2009 BUT system for NIST 2008 speaker recognition evaluation. In *Proc. Interspeech*, pp. 2335–2338

Campbell W, Gleason T, Navratil J, Reynolds D, Shen W, Singer E, Torres-Carrasquillo P 2006 Advanced language recognition using cepstra and phonotactics: MITLL system performance on the NIST 2005 language recognition evaluation. In *IEEE Odyssey Speaker and Language Recognition Workshop*

Cetin Ö, Magimai-Doss M, Livescu K, Kandtor A, King S, Bartels C, Frankel J 2007 Monolingual and crosslingual comparison of tandem features dervied from articulatory and phone MLPs. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 36–41

Chen X-X, Li A-J, Sun G-H, Hua W, Yin Z-G 2000 An application of SAMPA-C for standard Chinese. In *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, pp. 652–655, Beijing, China

Constantinescu A, Chollet G 1997 On cross-language experiments and data driven units for ALSP (automatic language independent speech processing). In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 606–613

Dalsgaard P, Andersen O 1992 Identification of mono- and poly-phonemes using acoustic-phonetic features derived by self-organizing neural network. In *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, pp. 547–550

Dines J, Saheer L, Liang H 2009 Speech recognition with speech synthesis models by marginalising over decision tree leaves. In *Proc. Interspeech*, pp. 1395–1398, Brighton, UK

Dines J, Yamagishi J, King S 2010 Measuring the gap between HMM-based ASR and TTS. *IEEE Special Topics Signal Process* 4(6): 1046–1058

Dutoit T, Pagel V, Pierret N, Bataille F, van der Vrecken O 1996 The MBROLA project: Towards a set of high quality speech synthesizers free of use for non-commercial purposes. In *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, pp. 1393–1396, Philadelphia, USA

Fosler-Lussier E, Morris J 2008 CRANDEM Systems: Conditional Random Field Acoustic Models for Hidden Markov Models. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 4049–4052

Fousek P, Lamel L, Gauvain J L 2008 Transcribing Broadcast Data Using MLP Features. In *Proc. Interspeech*, pp. 1433–1436

Fügen C, Stüker S, Soltau H, Metze F, Schultz T 2003 Efficient handling of multilingual language models. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 441–446

Gales M, Young S 2007 The application of hidden Markov models in speech recognition. *Foundations and Trends in Signal Processing* 1(3): 195–304

Gibson M 2009 Two-pass decision tree construction for unsupervised adaptation of HMM-based synthesis models. In *Proc. Interspeech*, pp. 1791–1794, Brighton, UK

Hermansky H, Fousek P 2005 Multi-resolution RASTA filtering for TANDEM-based ASR. In *Proc. Interspeech*, pp. 361–364

Hermansky H, Ellis D P W, Sharma S 2000 Tandem connectionist feature extraction for conventional HMM systems. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 1635–1638, Istanbul, Turkey

Hieronymus J L 1993 ASCII phonetic symbols for the world's languages: Worldbet. Technical Memo. 23, AT&T Bell Laboratories, Murray Hill, NJ 07974, USA

Hillard D, Hwang M, Harper M, Ostendorf M 2008 Parsing-based objective functions for speech recognition in translation applications. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process (ICASSP)*, pp. 5109–5112

Höge H, Draxler C, van den Heuvel H, Johansen F T, Sanders E, Tropf H S 1999 Speechdat multilingual speech databases for teleservices: across the finish line. In *Proc. European Conf. Speech Commun. Technol. (EUROSPEECH)*, pp. 2699–2702

Ikbal S 2004 *Nonlinear feature transformations for noise robust speech recognition*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Ikbal S, Misra H, Sivadas S, Hermansky H, Bourlard H 2004 Entropy based combination of tandem representations for noise robust ASR. In *Proc. INTERSPEECH-ICSLP-04*, pp. 2553–2556

International Phonetic Association (IPA) 1999 *Handbook of the International Phonetic Association. A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, The Edinburgh Building, Cambridge CB2 8RU, UK. ISBN-10:0521637511, ISBN-13:978-0521637510

Imseng D, Magimai-Doss M, Bourlard H 2010 Hierarchical multilayer perceptron based language identification. Idiap-RR Idiap-Internal-RR-104-2010, Idiap, May 2010. URL http://www.idiap.ch/∼dimseng/Idiap_IIR_104-2010.pdf

Kawahara H, Estill J, Fujimura O 2001 Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In *Proc. MAVEBA*, Florence, Italy

Khudanpur S P 2006 Multilingual language modeling. In (eds) T Schultz, K Kirchoff, *Multilingual speech processing*, chapter 6, pp. 169–205. Academic Press, USA

Kim W, Khudanpur S 2002 Using cross-language cues for story-specific language modelling. In *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, pp. 513–516

Kim W, Khudanpur S 2003 Cross-lingual lexical triggers in statistical language modeling. In *Proc. Empirical Methods in Natural Language Processing (EMNLP)*, pp. 17–24

Kim W, Khudanpur S 2004 Cross-lingual latent semantic analysis for language modeling. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. I–257–I–260

King S, Tokuda K, Zen H, Yamagishi J 2008 Unsupervised adaptation for HMM-based speech synthesis. In *Proc. of Interspeech*, pp. 1869–1872

Köhler J 1996 Multilingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. In *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, pp. 2195–2198

Köhler J 1998 Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 417–420

Köhler J 1999 Comparing three methods for multilingual phone models for vocabulary independent speech recognition tasks. In *Proc. of the ESCA-NATO Tutorial Workshop on Multilingual Interportability in Speech Technology*, pp. 79–84

Koishida K, Hirabayashi G, Tokuda K, Kobayashi T 1994 Mel-generalized cepstral analysis — a unified approach to speech spectral estimation. In *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, vol. 3, pp. 1043–1046, Yokohama, Japan

Kominek J 2009 *TTS From zero: Building synthetic voices for new languages*. PhD thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, USA

Kubala F, Bellegarda J, Cohen J, Pallett D, Paul D, Phillips M, Rajasekaran R, Richardson F, Riley M, Rosenfeld R, Roth B, Weintraub M 1994 The hub and spoke paradigm for CSR evaluation. In *Human*

Language Technology Conference: Proceedings of the Workshop on Human Language Technology, pp. 37–42, Plainsboro, NJ

Lamel L, Gauvain J-L, Adda G 2002 Lightly supervised and unsupervised acoustic model training. *Comput. Speech Lang.* 16: 115–129

Lamel L F, Gauvain J-L 1993 Cross-lingual experiments with phone recognition. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 2, pp. 507–510

Latorre J 2006 *A study on speaker adaptable speech synthesis*. PhD thesis, Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Tokyo, Japan

Lee L, Rose R 1998 A frequency warping approach to speaker normalisation. *IEEE Trans. Speech Audio Process.* 6: 49–60

Li X, Bilmes J 2006 Regularized adaptation of discriminative classifiers. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. I–237–I–240

Liang H, Dines J 2010 An analysis of language mismatch in HMM state mapping-based cross-lingual speaker adaptation. In *Proc. Interspeech*, pp. 622–625, Makuhari, Japan

Liang H, Dines J, Saheer L 2010 A comparison of supervised and unsupervised cross-lingual speaker adaptation approaches for HMM-based speech synthesis. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 4598–4601, Dallas, USA

Lööf J, Gollan C, Ney H 2009 Cross-language bootstrapping for unsupervised acoustic model training: Rapid development of a Polish speech recognition system. In *Proc. Interspeech*, pp. 88–91, Brighton, UK

Maskey S R, Black A W, Tomokiyo L M 2004 Bootstrapping phonetic lexicons for new languages. In *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, pp. 69–72, Jeju Island, Korea

McDonough J W 2000 *Speaker compensation with all-pass transforms*. PhD thesis, Johns Hopkins University

Misra H, Bourlard H, Tyagi V 2003 New entropy based combination rules in HMM/ANN multi-stream ASR. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. II–741–II–744

Morgan N, Zhu Q, Stolcke A, Sonmez K, Sivadas S, Shinozaki T, Ostendorf M, Jain P, Hermansky H, Ellis D, Doddington G, Chen B, Cetin O, Bourlard H, Athineos M 2005 Pushing the Envelope–Aside. *IEEE Signal Process. Mag.* 22(5): 81–88

Motlicek P 2009 Automatic out-of-language detection based on confidence measures derived from LVCSR word and phone lattices. In *Proc. Interspeech*, pp. 1215–1218, Brighton, UK

Motlicek P, Valente F 2010 Application of out-of-language detection to spoken term detection. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 5098–5101, Dallas, USA

Muthusamy Y K, Cole R A, Oshika B T 1992 The OGI multi-language telephone speech corpus. In *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, pp. 895–898, Banff, Alberta, Canada

Navratil J 2001 Spoken language recognition – a step toward multilinguality in speech processing. *IEEE Trans. Audio Speech Lang. Process.* 9(6): 678–685

Navratil J 2006 Automatic language identification. In (eds) T Schultz, K Kirchoff, *Multilingual speech processing*, chapter 8, pp. 233–271. Academic Press, USA

Ostendorf M, Bulyko I 2002 The impact of speech recognition on speech synthesis. In *Proc. IEEE Workshop on Speech Synthesis*, pp. 99–106, Santa Monica, USA

Osterholtz L, Augustine C, McNair A, Saito I, Sloboda T, Tebelskis J, Waibel A, Woszczyna M 1992 Testing generality in JANUS: A multilingual speech translation system. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 209–212

Paul D B, Baker J 1992 The design for the wall street journal-based CSR corpus. In *Human Language Technology Conference: Proceedings of the Workshop on Speech and Natural Language*, pp. 357–362, Harriman, NY

Pinto J, Yegnanarayana B, Hermansky H, Magimai-Doss M 2008 Exploiting contextual information for improved phoneme recognition. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 4449–4452

Pinto J, Sivaram G S V S, Magimai-Doss M, Hermansky H, Bourlard H 2011 Analysis of MLP based hierarchical phoneme posterior probability estimator. *IEEE Trans. Audio Speech Lang. Process.* 19(2): 225–241

Pinto J P 2010 *Multilayer Perceptron Based Hierarchical Acoustic Modeling for Automatic Speech Recognition*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Rabiner L R 1989 A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77(2): 257–286

Saheer L, Dines J, Garner P N, Liang H 2010a Implementation of VTLN for statistical speech synthesis. In *Proc. 7th Speech Synthesis Workshop*, Kyoto, Japan

Saheer L, Garner P N, Dines J 2010b Study of Jacobian normalisation for VTLN. Idiap-RR Idiap-RR-25-2010, Idiap Research Institute, Martigny, Switzerland

Saheer L, Garner P N, Dines J, Liang H 2010c VTLN adaptation for statistical speech synthesis. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 4838–4841, Dallas, USA

Schlüter R, Macherey W, M"uller B, Ney H 2001 Comparison of discriminative training criteria and optimization methods for speech recognition. *Speech Commun.* 34(3): 287–310

Schultz T 2006 Multilingual acoustical modeling. In (eds) T Schultz, K Kirchoff, *Multilingual Speech Processing*, chapter 4, pp. 71–122. Academic Press, USA

Schultz T, Waibel A 1997a Fast bootstrapping of LVCSR systems with multilingual phoneme sets. In *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, vol. 1, pp. 371–374, Rhodes, Greece

Schultz T, Waibel A 1997b Fast bootstrapping of LVCSR systems with multilingual phoneme sets. In *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 371–374

Schultz T, Waibel A 1998 Language independent and language adaptive large vocabulary speech recognition. In *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, pp. 1819–1822

Schultz T, Waibel A 2001 Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Commun.* 35(1–2): 31–50

Schultz T, Rogina I, Waibel A 1996 LVCSR-based language identification. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 2, pp. 781–784

Siemund R, Höge H, Kunzmann S, Marasek K 2000 SPEECON – speech data for consumer devices. In *Proc. 2nd Int. Conf. Language Resources & Evaluation*, pp. 883–886, Athens, Greece

Silén H, Hel E, Nurminen J, Gabbouj M 2009 Parameterization of vocal fry in HMM-based speech synthesis. In *Proc. Interspeech*, pp. 1775–1778, Brighton, UK

Sivadas S, Hermansky H 2004 On Use of Task Independent Training Data in Tandem Feature Extraction. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. I–541–I–544

Sproat R (ed) 1997 *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Kluwer Academic Publishers, Norwell, Massachussetts, USA

Stolcke A, Grézl 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. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. I–321–I–324

Suendermann D, Hoege H, Bonafonte A, Ney H, Hirschberg J 2006 TC-Star: Cross-language voice conversion revisited. In *Proc. TC-Star Workshop*, Barcelona, Spain

Sugiyama M 1991 Automatic language recognition using acoustic features. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 813–816

Tokuda K, Masuko T, Miyazaki N, Kobayashi T 2002 Multi-space probability distribution HMM. *IEICE Trans. Inf. Syst.* E85-D(3): 455–464

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*, pp. 2695–2698

Traber C, Huber K, Nedir K, Pfister B, Keller E, Zellner B 1999 From multilingual to polyglot speech synthesis. In *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 835–838, Budapest, Hungary

Valente F 2009 A novel criterion for classifiers combination in multistream speech recognition. *IEEE Signal Process. Lett.* 16(7): 561–564

Valente F, Hermansky H 2008 Hierarchical and Parallel Processing of Modulation Spectrum for ASR applications. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 4165–4168

Valente F, Magimai-Doss M, Plahl C, Ravuri S 2009 Hierarchical Modulation spectrum for the GALE project. In *Proc. Interspeech*, pp. 2963–2966

Valente F, Magimai-Doss M, Plahl C, Ravuri S, Wang W 2010 A comparative large scale study of MLP features for Mandarin ASR. In *Proc. Interspeech*, pp. 2630–2633

Wan V, Hain T 2006 Strategies for language model data collection. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. I–1069–I–1072, Toulouse, France

Weng F, Bratt H, Neumeyer L, Stolcke A 1997 A study of multilingual speech recognition. In *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 359–362

Wester M 2010a Cross-lingual talker discrimination. In *Proc. Interspeech*, pp. 1253–1256, Makuhari, Japan

Wester M 2010b The EMIME bilingual database. Technical Report EDI-INF-RR-1388, The University of Edinburgh, UK

Wester M *et al* 2010 Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project. In *Proc. 7th Speech Synthesis Workshop*, Kyoto, Japan

Wheatley B, Kondo K, Anderson W, Muthuswamy Y 1994 An evaluation of cross-language adaptation for rapid HMM development in a new language. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. I/237–I/240

Wu Y-J, Wang R-H 2006 Minimum generation error training for HMM-based speech synthesis. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. I–89–I–92, Toulouse, France

Wu Y-J, King S, Tokuda K 2008 Cross-lingual speaker adaptaton for HMM-based speech synthesis. In *Proc. Interspeech*, pp. 528–531, Brisbane, Australia

Wu Y-J, Nankaku Y, Tokuda K 2009 State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis. In *Proc. Interspeech*, pp. 528–531, Brighton, UK

Yamagishi J, Lincoln M, King S, Dines J, Gibson M, Tian J, Guan Y 2009 Analysis of unsupervised and noise-robust speaker-adaptive HMM-based speech synthesis systems toward a unified ASR and TTS framework. In *Proc. Blizzard Challenge Workshop*, Edinburgh, UK

Zen H, Tokuda K, Masuko T, Kobayashi T, Kitamura T 2007 A hidden semi-Markov model-based speech synthesis system. *IEICE Trans. Inf. Syst.* E90-D(5): 825–834

Zen H, Tokuda K, Black A W 2009 Statistical parametric speech synthesis. *Speech Commun.* 51(11): 1039–1064

Zheng J, Cetin O, Hwang M-Y, Lei X, Stolcke A, Morgan N 2007 Combining discriminative feature, transform, and model training for large vocabulary speech recognition. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. IV–633–IV–636

Zhu Q, Chen B, Morgan N, Stolcke A 2004 On using MLP features. In *Proc. INTERSPEECH-ICSLP-04*, pp. 921–924

Zissman M A 1996 Comparison of four approaches to automatic language identification of telephone speech. *IEEE Trans. Audio Speech Lang. Process.* 4(1): 31–42

Zissman M A, Berkling K M 2001 Automatic language identification. *Speech Commun.* 35: 115–124