SPEAKER- AND LANGUAGE-INDEPENDENT SPEECH RECOGNITION IN MOBILE COMMUNICATION SYSTEMS

Olli Viikki, Imre Kiss, Jilei Tian

Nokia Research Center, Speech and Audio Systems Laboratory, Tampere, Finland
Email: {olli.viikki,imre.kiss,jilei.tian}@nokia.com

ABSTRACT

In this paper, we investigate the technical challenges that are faced when making a transition from the speaker-dependent to speaker-independent speech recognition technology in mobile communication devices. Due to globalization as well as the international nature of the markets and the future applications, speaker independence implies the development and use of language-independent ASR to avoid logistic difficulties. We propose here an architecture for embedded multilingual speech recognition systems. Multilingual acoustic modeling, automatic language identification, and on-line pronunciation modeling are the key features which enable the creation of truly language- and speaker-independent ASR applications with dynamic vocabularies and sparse implementation resources. Our experimental results confirm the viability of the proposed architecture. While the use of multilingual acoustic models degrades the recognition rates only marginally, a recognition accuracy decrease of approximately 4% is observed due to sub-optimal on-line text-to-phoneme mapping and automatic language identification. This performance loss can nevertheless be compensated by applying acoustic model adaptation techniques.

1. INTRODUCTION

Over the past few years, Automatic Speech Recognition (ASR) technology has hit mobile phones. Speaker-trained name dialer is today probably one of the most widely distributed ASR applications in the world. Despite the fact that more advanced speech recognition applications have already been introduced for other platforms during several years, there are many aspects which justify the use of this simple technology in portable mobile communication devices. Multilinguality, low complexity implementation, and a high degree of robustness against background noise are the key qualities of the speaker-dependent ASR technology.

The undoubted advantage of speaker-trained technology is its inherent support to various languages as all users train the recognition system to match their language and pronunciation characteristics. Multilinguality is one of the main requirements set for speech recognition applications running on mobile devices. A truly multilingual speech recognition system can simultaneously support several languages and is able to cope with non-native speakers, dialects, accents, and multilingual vocabulary items. There are also substantial financial facts which justify the development of multilingual speech recognition systems. For such products, which are sold world-wide, it is utterly important that there is no need to develop different versions of the same product for different language regions. Despite its indisputable importance, surprisingly little efforts have been put on multilingual ASR research compared for instance with noise robust speech recognition. In addition to multilinguality, noise robustness and low complexity implementation are also required for all applications realized on mobile devices due to the wide range of different operating conditions and sparse implementation resources. An isolated word recognition task combined with the speaker-dependent ASR technology guarantees high enough recognition rates on various languages across various types of operating environments. It is apparent that the same performance and implementation requirements must also be met when shifting from speaker-dependent to speaker-independent ASR technology.

Thanks to the progress done in technology development and the availability of more powerful implementation platforms, it is obvious that speech recognition in mobile devices also trends towards speaker independence. However, it is still unrealistic to expect that this technology shift would mean the possibility to run very advanced speech recognition applications on these very resource limited platforms.

In this paper, we discuss about technical solutions needed for enabling the use of language- and speaker-independent speech recognition technology in embedded systems, e.g. mobile phones. The focus is still on applications with isolated word, or very restricted continuous, speech input. The remainder of the paper is organized as follows. Section 2 discusses the characteristics of mobile communication systems and their implications on ASR. In Section 3, a speech recognition architecture for multilingual systems is described. Section 4 summarizes some experimental results obtained in the recognition tests to verify the viability of the chosen technical solutions.

2. ASR IN MOBILE COMMUNICATION DEVICES

It is generally acknowledged that there is enormously potential in speech recognition to renew the ways how the users will interact with various communication devices in the future. There are, however, several constraints that need to be considered when starting to integrate voice control functionality in products which are used by hundreds of millions of people world-wide. In this section, we address some of the most important requirements that need to be considered until ASR can be an integral part of the user interface of global products.

2.1 Usability Requirements

It is obvious that the current wide use of speaker-dependent ASR technology is mainly due to technical limitations, as from the user's perspective, the training process is often seen as an additional bur-

den. While it may sometimes be acceptable to let the users train/adapt the recognizer to their voice in an enrollment session, in such products where speech recognition is not the main feature, this cumbersome and lengthy training process should be avoided if possible. Since the users are only seldom willing to train many vocabulary items, the vocabulary size in speaker-dependent ASR applications is often very limited. These problems can be alleviated using the speaker-independent ASR technology, as it improves the ease-of-use and provides a wider range of application possibilities than when relying on the use of speaker-dependent technology.

2.2 Implementation and Application Needs

Compared with a normal PC environment, the implementation resources in embedded systems are sparse both in terms of processing power and memory. Because the factory price has a crucial importance in mass-produced products, it is important to pay attention to minimize all implementation costs. A compact implementation of the ASR engine can result in substantial cost savings making the product more competitive in terms of price.

There are many ASR application possibilities in mobile communication systems. Both the vocabulary size and the type of speech input, i.e. isolated words vs. continuous speech, can change application by application. Since the vocabulary is dynamic in the most of ASR applications, it is clear that acoustic modeling cannot be based on whole-word models, but smaller sub-word based acoustic units are needed. There are many benefits supporting the use of sub-word based acoustic modeling. We can realize easily portable ASR applications for different recognition tasks. It is also possible to let the users modify the vocabulary items according to their own needs and preferences. Finally, considerable development cost reductions can be achieved as the expensive and time-consuming application-specific data collection can be avoided.

The use of sub-word models requires a pronunciation modeling scheme to define how different sub-word units are concatenated to words. Due to their large memory requirements, extensive pronunciation lexicons, which are commonly used in PC based ASR systems, cannot though be used in embedded systems. *On-line* techniques[1] are instead required to make a conversion between the written and spoken language. As the vocabulary items are not always monolingual, pronunciation modeling needs to be designed such that it can also cope with multilingual vocabulary words. Automatic text based language identification is therefore required to choose the valid pronunciation scheme for different multilingual vocabulary items.

2.3 Recognition Performance

Noise robustness and immunity to speaker variability are probably the two most important requirements that are common to all recognition systems. Although noise robustness has intensively been studied for the last decade, it remains one of the key challenges in speech recognition. As mobile devices are used virtually everywhere, a high degree of noise robustness is an obligatory requirement. This requirement also restricts the type of applications that can be included in portable products. More progress is required in noise robust ASR research until for example large vocabulary continuous speech recognition can successfully be utilized in mobile communication systems.

In addition to noise robustness, the ASR system must also cope with speaker variability. Due to the multilingual acoustic models, it is obvious that the mismatch between acoustic models and the speaker's speech is greater in multi- than in monolingual systems.

3. MULTILINGUAL ASR ARCHITECTURE

Figure 1 illustrates the proposed architecture for multilingual ASR systems. The multilingual ASR engine consists of three key units: automatic language identification, on-line pronunciation modeling, and multilingual acoustic modeling modules. The assumption is that vocabulary items are given in the textual form. First, the Language Identification (LID) module detects the language of the vocabulary item. Once this has been determined, an appropriate on-line pronunciation modeling scheme is applied to get the phoneme sequence associated with the written form of the vocabulary item. Finally, the recognition model for each vocabulary item is constructed by concatenating the multilingual acoustic models. With these basic modules, the recognizer can automatically cope with multilingual vocabulary items without the user assistance. In the remainder of this section, these basic building blocks of the multilingual ASR system are investigated in greater details.

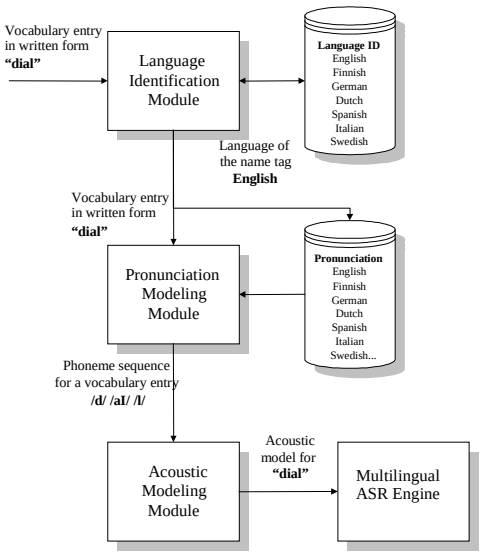

Figure 1: Architecture for a multilingual ASR system.

3.1 Multilingual Acoustic Modeling

The performance of any ASR system is highly dependent on the quality of the acoustic models. When aiming at supporting multiple languages and having restricted memory capabilities, it is obvious that one must make compromises in the modeling accuracy. The sufficiency of memory is the main problem in acoustic modeling. Therefore, some of the most widely used acoustic modeling schemes cannot be utilized in embedded systems. It is easy to understand that commonly used context-dependent acoustic models are not an attractive solution due to their large memory requirements. Language-dependent acoustic models are also problematic,

[1] This is of course not needed if the application has a fixed vocabulary when the system designer can specify the pronunciation for each vocabulary item.

particularly, if we need to support several languages at the same time.

To have a reasonable number of acoustic models, we selected monophone HMMs as a basic acoustic building block. The monophone models are further shared across different languages and the parameters of continuous density monophone HMMs are trained on multilingual speech corpora for having as small number of models as possible. We chose the International Phonetic Alphabet (IPA) [1] to define the phoneme inventory for the multilingual ASR engine. Some language-specific modifications have though been included in the IPA phoneme set either to further reduce the number of models or to increase the modeling accuracy. In addition to the small number of acoustic models, the multilingual approach also makes it possible to support languages for which there is no speech data available for HMM parameter estimation. As shown in Section 4, an acceptable recognition accuracy can be achieved for an unseen language by defining only the valid pronunciation of vocabulary items.

Acoustic model adaptation has been found to be an efficient method to increase the speaker-specific recognition rate by several researchers [2]. Since multilingual acoustic models cannot characterize the language-specific details as accurately as their monolingual counterparts, the importance of model adaptation is even greater in multi- than in monolingual ASR systems. Besides improving speaker-specific recognition rate, it is also possible to increase the performance for unseen languages that are only supported at the pronunciation modeling level. This enables us to support minority languages for which no large enough speech corpora exist. After the user has uttered a few utterances, the adapted acoustic models provide a recognition rate that is comparable to those languages that have been seen in the training phase.

As mentioned in Section 2.1, off-line adaptation is only seldom an acceptable solution, and therefore, the adaptation process often needs to be made transparent to the user. Another advantage of on-line adaptation is that it is capable of adapting the system to certain operating conditions preferred by the user.

3.2 Automatic Pronunciation Modeling

On-line pronunciation modeling, i.e. Text-to-Phoneme (T2P) mapping, is an obligatory feature in embedded systems with dynamic vocabularies where it is not feasible to have large pronunciation dictionaries for several languages. If the pronunciation of a language is very regular, e.g. in Finnish or Japanese, the T2P mapping module is very compact as it can be realized from a finite set of rules. There are, however, many languages, English being the best example, whose pronunciation cannot accurately be expressed using a rule set. To gain a high performance T2P mapping for irregular languages, it is necessary to have large text resources.

Decision trees have successfully been used to compress large pronunciation dictionaries [4][6]. The T2P irregularity of the language controls the size and accuracy of the decision tree based pronunciation model. If the number of T2P exceptions is small, the decision trees do not become very big. However, the size of the decision tree based T2P model increases rapidly if there are many pronunciation exceptions in the language. T2P mapping can also be implemented using neural nets [3] when the module becomes very compact.

3.3 Language Identification

The task of the Language Identification (LID) module is to identify the language of each vocabulary item based on its textual form. This decision is utilized to choose an appropriate text-to-phoneme mapping technique for each vocabulary item. Since the result of the LID module is not always unambiguous, it is important to provide multiple results and pronunciations for certain vocabulary items.

In general, a text based LID is a fairly new research topic. A straightforward approach is to utilize the occurrence probabilities of different letter combinations and certain language-specific letters [5]. The drawback of this *n*-gram modeling is that the size of the LID module increases rapidly with the higher values of *n*.

4. EXPERIMENTAL RESULTS

The objective of the performance evaluation was to confirm the technical viability of the proposed multilingual speech recognition architecture, i.e. how much the recognition rate is affected by the approximations made in acoustic modeling, language identification, and pronunciation modeling. A multilingual ASR engine supporting five European languages, English, German, Spanish, Finnish, and Italian, was created according to Figure 1. The amount of the multilingual acoustic training data was approximately balanced in terms of different languages. However, no acoustic training data was available for Italian. Small modifications to the IPA phoneme definitions were made to reduce the number of acoustic models, e.g. no separate models were trained for double consonants and vowels in Finnish. The phoneme set for Italian was constructed from the phonemes occurring in the other four languages.

4.1 Front-End, Acoustic Models, and Test Set-Up

A set of 12 MFCC coefficients and log-energy, together with their first- and second-order time derivatives, were extracted from a continuous-time speech signal sampled at 8 kHz. Three-state, left-to-right continuous density HMMs were trained to characterize all 66 monophones that were chosen to represent the spoken sounds of the five test languages.

The language-specific test vocabulary consisted of 120 isolated commands for each test language. There were both "native" and "non-native" items included in the vocabulary. The majority of vocabulary entries matched the language of the test speaker. Each command was repeated twice by all test speakers.

4.2 Multilingual Acoustic Modeling

First, it was tested how much the recognition rate degrades if we replace the monolingual acoustic models by the multilingual HMMs. For each language, except for Italian, the language-dependent acoustic models were estimated, and their performance was compared to the multilingual phoneme set. As shown in Table 1, the use of multilingual HMMs decreased the recognition rates only marginally, except in the case of Finnish, for which the rates improved considerably. It should be noted that neither on-line pronunciation modeling nor LID was used in these tests, but both

pronunciations and language identities of all vocabulary items were specified by a human expert.

Language	Monolingual HMMs		Multilingual HMMs	
	8 mix	16 mix	8 mix	16 mix
English	94.5	95.2	93.3	94.7
German	93.3	94.6	93.1	94.1
Spanish	94.6	95.4	94.5	95.6
Finnish	96.4	96.5	98.1	98.8
Average	94.7	95.4	94.8	95.8

Table 1: Recognition rate comparison between mono- and multilingual acoustic models.

4.3 On-line T2P and Automatic LID

While the results in Table 1 were obtained with the error-free pronunciations and language identification decisions, the goal of the following tests was to study how the recognition rate is affected when automatic methods were applied to these tasks. Decision tree based T2P modules were created from English, German, and Spanish pronunciation dictionaries. A rule set was defined for Finnish and Italian. An *n*-gram (*n*=2) based LID module was also created from large text resources. Figure 2 depicts the results obtained in these tests with the 8-mixture multilingual HMMs.

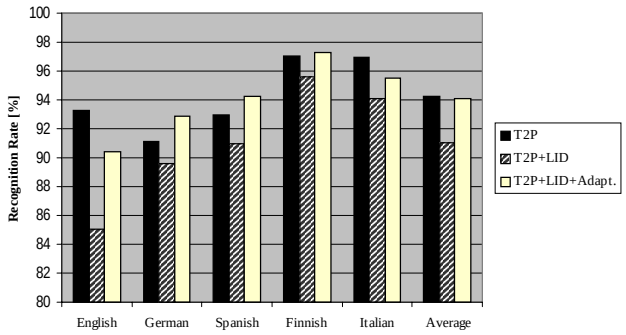

Figure 2: The effect of automatic T2P and LID on the recognition accuracy.

Not surprisingly, a small performance degradation is observed for all languages due to the sub-optimal pronunciations. The performance drop is nevertheless insignificant for all languages. It is also interesting to note that very high recognition rates were achieved for Italian although no Italian data was present during training. These results suggest that multilingual acoustic models can successfully be applied also for unseen languages for which only pronunciation information is available

Automatic LID appears to degrade the performance more than on-line pronunciation modeling. Particularly for English, the rates were affected quite drastically. The tests also indicate the importance of acoustic model adaptation to compensate the performance losses due to erroneous T2P and LID decisions.

4.4 On-line Adaptation Experiments

Inter-speaker variability (accents, dialects etc.), environmental mismatch between training and testing conditions, as well as the language mismatch between the multilingual acoustic models and the test language, are the three major sources resulting in performance degradation when using multilingual speaker-independent acoustic models. Figure 3 illustrates the recognition performance gain that was obtained when supervised on-line MAP adaptation of Gaussian means [2] was included in the recognition system. The tests were done both in clean and noisy test environments with on-line T2P. The noisy utterances were created by adding various types of noise signals (car noise, babble noise, music) to clean waveforms at the Signal-to-Noise Ratio range of +20…+5 dB. Adaptation results clearly show that acoustic model adaptation should be an integral part of the multilingual ASR system.

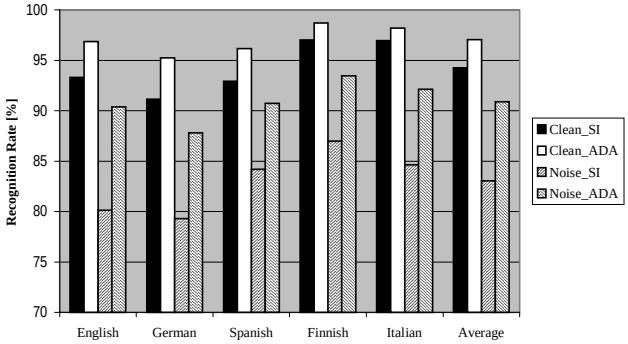

Figure 3: The performance comparison of the speaker-independent (SI) and on-line adapted (ADA) multilingual ASR systems both in clean and noisy conditions, 8-mixture HMMs.

5. CONCLUSIONS

In this paper, we have proposed a framework for multilingual speech recognition in mobile communication devices. Compared with a monolingual speech recognition architecture, this revised ASR framework includes three new modules, namely multilingual acoustic modeling, on-line T2P, and automatic LID. By relying on these new modules, it is feasible to realize a high performance multilingual ASR system on the resource sparse implementation platform that can deal with dynamic and multilingual vocabularies. Preliminary experimental results for five European languages show the usefulness of the proposed ASR architecture.

REFERENCES

[1] The International Phonetic Association, *Handbook of the International Phonetic Association (IPA)*, Cambridge University Press, Cambridge, UK, 1999.

[2] J. L. Gauvain, C.-H. Lee, "Maximum a Posteriori Estimation of Multivariate Gaussian Mixture Observations of Markov Chains", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 2, pp. 291-298, April 1994.

[3] K. J. Jensen, S. Riis, "Self-Organizing Letter Code-Book for Text-to-Phoneme Neural Network Model", *Proc. of ICSLP'00*, Beijing, China, 2000.

[4] V. Pagel, K. Lenzo, A. W. Black, "Letter to Sound Rules for Accented Lexicon Compression," *Proc. of ICSLP'98*, Sydney, Australia, 1998.

[5] J. Prager, "Linguini: Language Identification for Multilingual Documents", *32nd Hawaii International Conference on System Sciences*, pp. 1-11, Hawaii, 1999.

[6] J. Suontausta, J. Häkkinen, "Decision Tree Based Text-to-Phoneme Mapping For Speech Recognition", *Proc. of ICSLP'00*, Beijing, China, 2000.