ASR IN PORTABLE WIRELESS DEVICES

Olli Viikki

Nokia Research Center, Speech and Audio Systems Laboratory
Tampere, Finland
Email: olli.viikki@nokia.com

ABSTRACT

This paper discusses the applicability and role of automatic speech recognition in portable wireless devices. Due to the author's background, the viewpoints are somewhat biased to mobile phones, but many of the aspects are nevertheless common for other portable devices as well. While still dominated by the speaker-dependent technology, there are today signs that also in wireless devices, ASR trends towards speaker-independent systems. As these modern communication devices are usually intended for mass markets, the paper reviews the ASR areas that are relevant for speech recognition on low cost embedded systems. In particular, multilingual ASR, low complexity ASR algorithms and their implementation, and acoustic model adaptation techniques play a key role enabling cost effective realization of ASR systems. Low complexity and advanced noise robust ASR algorithms are sometimes conflicting concepts. The paper also briefly reviews some of the most important noise robust ASR techniques that are well suited for embedded systems.

1. INTRODUCTION

Automatic Speech Recognition (ASR) applications and services have already for some time attracted a significant interest in portable wireless devices, such as mobile phones or personal digital assistants (PDA). New voice controlled services, and in particular, the usability aspects have driven the introduction of the ASR technology in these systems. The small physical size of these handheld devices, together with the rich set of various applications, put very high requirements for input and output technologies to ensure the ease of use of these devices. This paper discusses the requirements and possible solutions that one needs to take into account when starting to integrate ASR technology into portable devices with low implementation resources.

Portability and low implementation resources are the two key characteristics of wireless devices that mainly determine the technical ASR solutions to be used in real-world systems. Furthermore, wireless devices, e.g. mobile phones, are typically targeted to be sold in large volumes at mass markets. It is therefore essential that ASR can be brought into practical use in a profitable way. The development and implementation of cost-effective ASR solutions, which is crucial in real-world systems, is an aspect that has up to now partially been ignored by the speech recognition research community. Even though fairly advanced speech recognition services and applications have been introduced during the past years, the simple and straightforward speaker-dependent (speaker-trained) ASR technology is still widely applied in embedded wireless devices. Today, only this simple technology is capable of meeting all these real-world requirements. Despite the progress made in microelectronics, the step from speaker-dependent to speaker-independent ASR requires further advances in noise robust, low complexity, and multilingual speech recognition.

In addition to pure terminal based ASR implementations, speech recognition services can also be provided remotely, through telecommunication networks. Due to the modular structure of ASR systems, it is possible to place the entire speech recognition service to a network server end, or the recognition system can also be distributed between the terminal and the server, as is done in the Distributed Speech Recognition (DSR) framework. The possibility to use efficient network servers in ASR application development enables the creation of more versatile and computationally complex applications than when relying on speech recognizers implemented on terminals only.

The remainder of this paper is organized as follows. First, the nature of ASR applications in portable wireless devices is discussed. Both application and implementation perspectives are reviewed. Next, in the more technical part of the paper, various issues related to ASR implementation on portable wireless devices are reviewed. In particular, the objective is to discuss the real-world issues that one is facing when trying to introduce advanced speech recognition technology for embedded portable systems. Finally, in Section 5, the paper is summarized with the brief discussion about the future of ASR in portable wireless devices.

2. SPEECH RECOGNITION IN MOBILE AND WIRELESS SYSTEMS

It is generally believed that speech recognition has lots of application potential in mobile communication devices and it can truly enhance the way in which these devices are going to be used in the future. There are two major facts that make this belief realistic:-

1. The development of easy-to-use user interfaces is crucial for mobile phones and other portable wireless devices. The small size and the continuously increasing number of new applications and functions challenge the conventional user interface solutions. Voice user interface has many good

qualities and it can be considered as one of the key technologies that enable the further miniaturization of devices.

2. Mobile communication systems have an enormous user population. It has been estimated that the mobile phone user base will exceed one billion users during the first half of 2002 [25]. This huge user base could greatly benefit from a convenient access to various network services and Internet.

2.1. Characteristics of Wireless ASR

Speech recognition in mobile devices has certain special requirements and characteristics that make it different from general speech recognition. Even if there have recently been some attempts to direct speech recognition research activities to meet the requirements of wireless ASR, many of the fundamental technical issues have only been addressed very superficially. In the following, some basic properties of ASR applications running on mobile terminals are summarized.

- *Operating environment* can range from clean to very noisy conditions. Not only handsets are used, but the use of car hands-free and headset devices is today customary.
- Both *processing power and memory resources* limited. As the amount of memory is a major cost factor in real-world systems, its use is carefully optimized by setting severe limitations to the ASR technology. Speech recognition competes for these small resources with several other resource-hungry technologies, e.g. imaging, video, and entertainment audio.
- Portable devices are powered by batteries which means that *computing resources* are available only for short periods of time. Usually, if the device is not actively used by any application, the system is in the stand-by mode for maximizing the battery life-time. Due to this, it is not currently feasible to run "continuous listening" type of ASR applications [7], or perform any background computation (e.g. online background noise estimation) if no power supply is available.
- Mobile handsets are intended for mass markets so that they are very *cost critical* appliances. All functions and new features that increase the costs per unit need to be well justified.
- Mobile phones are *global* devices. The GSM technology is used for instance in more than 100 countries worldwide. Although it would be technically feasible to create country- or language-specific devices, the logistic and cost issues make this in practice not possible. A new ASR framework is required to manage different languages in the cost efficient way.
- Mobile phones are today considered highly *personal* devices. The personalization of terminals, e.g. own ringing tones and various display icons, is nowadays very popular. This suggests that various ASR personalization techniques, e.g. speaker adaptation, can be utilized for maximizing the speaker-specific recognition rate.
- Terminals with the *network capabilities* open new possibilities to update speech recognizers, acoustic models, language configurations, or almost any part of a recognition system.

2.2. Towards Speaker Independence

As of today, speaker-dependent speech recognition is the major technology in wireless devices, i.e. mainly in mobile phones. There are several good qualities behind the popularity of this simple technology. Since recognizers are trained by the users, the systems inherently support all languages without any additional efforts. Moreover, the high degree of noise robustness combined with the low implementation complexity satisfy the requirements set for the recognition accuracy and implementation costs.

Even though speaker-trained technology appears to be a technically proven solution for embedded wireless devices, there are several application-specific issues and usability reasons why speaker-independent ASR technology would be preferred. There is only a narrow range of different applications which can be implemented using a speaker-dependent recognition engine. It is apparent that future visions cannot be realized if one has to rely on a technology that requires the users to go through a tedious and time consuming training step for all vocabulary items to be recognized. As speaker-independent ASR systems have already been available for PCs and network based ASR services, speaker-independent technology is also a logical step in embedded devices to enable the realization of more versatile ASR applications.

2.3. ASR in Consumer Products

When speech recognition is included in any kind of real-world appliances, whether mobile phones or coffee machines, the use of speech control must always be made as transparent as possible for the users. Speech recognition engineers need to remember that foremost, the users are buying a device, not a speech recognition system. In addition to speech control, other criteria, e.g. what other applications can be found on the device, the design, the usability, the "cool" factor, etc., may often play the greater role than ASR when the user is judging between different technical solutions.

Since ASR is still today very much an imperfect technology, it is essential to apply all possible techniques which are capable of maximizing the recognition performance. One should, however, be careful not to over-emphasize the importance of speech recognition in the host-application, but the user's perspective needs primarily to be taken into account when these algorithms are implemented.

While the users may be motivated to carry out a long enrollment session in dictation systems, speaker adaptation and other similar techniques must be hidden from the user in the domain of consumer appliances. In the same way, the use of speech recognition should not be dependent on the use of some special hardware. In noise robust speech recognition, special types of microphones are for example often used to maximize the Signal-to-Noise Ratio (SNR) of the input signal. Head-mounted microphones or microphone arrays should only be used if they are a normal part of an appliance. If the user is requested to do special hardware installations just for the sake of ASR, the speech control feature will not gain the wide acceptance of large user groups, but only some of the most technically oriented end-users are likely to use the feature.

2.4. Network Speech Recognition

Network ASR for mobile phones is in essence the same as interactive voice services that have already for some time been provided for fixed telecommunication networks. Compared with fixed networks, low bit rate speech coding methods used in mobile networks and the high degree variability of radio channels tend to distort the original speech signal resulting in recognition performance degradations [5]. The Distributed Speech Recognition (DSR) framework [11], where instead of the speech signal, only the output of the front-end module is sent to the server, has been introduced to alleviate these problems typical for mobile communication.

Mobile networks can also open some new application areas, e.g. location based services, which are not as relevant for the users of fixed networks. Mobile networks are also capable of combining speech and data transmission which enables the creation of multimodal applications where several input and output modalities can simultaneously be utilized.

Due to their different technical properties, terminal and network based ASR are best suited for different types of applications. Terminal based ASR is considered as a value-added technology which enriches the feature set included in the products. ASR is typically used to provide spoken shortcuts to various UI functions as done for instance in name dialing.

Network ASR is more directed towards service type of ASR applications. Thanks to large implementation resources, network ASR enables the creation of very sophisticated ASR services. These systems can also rely on large application-specific databases, e.g. timetable and stock quote inquiry applications. Due to various delays associated with communication between the terminal and server, network ASR is not though ideal for applications that require a strict real-time throughput. No terminal critical functions should be put behind the network ASR system, as the availability of the recognizer may be limited because of the lack of network connection, or the number of too many simultaneous users. From the user's perspective, it is important to note that the use of network ASR applications is not necessarily free of charge, as one may need to pay for a network connection. To be successful, network ASR applications must therefore be such that they attract users even though the services are not offered for free. Either the content behind the application must be interesting, or alternatively, speech should be clearly the best (only?) way to have an access to the service.

Even though terminal and network speech recognition are sometimes in certain business contexts considered as the competing approaches, they are technically very much complementary methods. Together, they provide a more enriched set of ASR applications and services for wireless devices, and therefore, it obvious that both of these two implementation options will co-exist in the future.

3. TOWARDS COST-EFFICIENT REALIZATION OF ASR SYSTEMS

Wireless devices, e.g. mobile phones, are typically aimed to be sold in high volumes in mass markets. This fact also needs to be considered when making decisions on the speech recognition technology. Low complexity implementation, both in terms of memory and computational overhead, is essential for minimizing the hardware and manufacturing costs. Because the factory price has a crucial importance in mass-produced products, it is necessary to minimize all implementation costs. A compact implementation of the ASR system can result in substantial cost savings making the product more competitive in terms of retail price.

In the general level, cost efficient implementation of speech recognition can be separated into two parts: 1) the efficient and compact implementation of ASR systems and 2) the minimization of development costs of an ASR system.

3.1. Low Complexity ASR

When discussing low complexity ASR, one should distinguish between two different concepts:-

- Low complexity implementation of ASR algorithms
- Low complexity ASR algorithms

While the first concept refers to the efficient implementation of the "standard" ASR algorithms, e.g. good practices to realize the standard Viterbi search, the second bullet covers a set of advanced methods that are capable of providing the same recognition performance as the baseline system with substantially smaller resource requirements. Obviously, both of these aspects must be well mastered in order to implement the recognizer on an embedded platform. In this paper, the emphasis is on the second bullet.

The decoding speed and the memory consumption are two aspects which are usually considered when evaluating the complexity of an ASR system. These two aspects are usually dependent on each other, i.e. by using more memory it is possible to speed up the decoding process, and vice versa. Due to the nature of ASR applications running on portable terminals (see Section 2), the high memory consumption and related implementation costs are today a greater issue than the decoding speed. Therefore, the following discussion focuses on the methods that are capable of reducing the memory consumption even if they have also turned out to be fairly efficient in terms of decoding speed. Furthermore, the discussion is limited to command and control type of applications which do not require any language models to be used in recognition.

3.1.1. Memory Efficiency

Both in speaker-dependent and –independent ASR systems the major memory use is due to acoustic models. Continuous Density Hidden Markov Models (CDHMM) are today almost a de facto acoustic modeling technique used in ASR. The accuracy of CDHMMs is often improved by increasing the number of Gaussian distributions within HMM states. The increase of Gaussian mixtures indeed improves the recognition accuracy, but this performance gain is achieved at the expense of memory and computational complexity. The challenge in memory efficient acoustic modeling is then to find a more compact representation of acoustic characteristics of speech than that of CDHMMs without degrading the recognition rate.

Substantial memory reduction was achieved when Gaussian distributions were replaced by artificial neural nets [13]. Compared with continuous density HMMs, the artificial neural net based HMMs were able to provide approximately the same recognition performance with 3-6 times less parameters.

Efficient acoustic model representation can also be achieved by quantizing the parameters of CDHMMs. A quantized parameter HMM framework (qHMM) was introduced in [19][20] where it is observed that efficient quantization of Gaussian mixtures can at least half the memory consumption without sacrificing the recognition accuracy. If small performance degradations are acceptable, it is possible to reduce the memory consumption up to 70%.

It should also be noted that acoustic model adaptation techniques that have commonly been applied to speaker and environment adaptation with very impressive results, enable the use of more compact-sized models as well. By applying the basic adaptation techniques developed for CDHMMs, e.g. MAP [1] and MLLR [9] adaptation, it is also possible to reduce the memory footprint without a loss of performance with respect to a non-adapted ASR system.

In addition to minimizing the memory occupied by acoustic models, it is also necessary to minimize the run-time memory consumption. Run-time memory consumption is typically directly proportional to the vocabulary size and the complexity of grammar networks. Tree-structured grammars, Viterbi beam search [17], and grammar minimization using finite-state transducers [10] have widely been found effective to reduce the use of run-time memory while still maintaining the recognition rate.

3.2. Cost Efficiency Through Multilingual ASR

Language dependency is an inherent problem associated with the speaker-independent ASR technology. One important, but often neglected, driver behind multilingual ASR systems is arguably the need to develop the technology in the cost efficient way. The development of speaker-independent ASR systems is both time-consuming and expensive. Large data collection procedures followed by acoustic model training steps involve a great deal of time and man-power. If the full development cycle always needs to take place for all languages to be supported, the cost of speaker-independent ASR becomes very high making it difficult to integrate this technology into practical systems.

To reduce the development costs, it is therefore important to have a flexible multilingual ASR framework to which it is easy to add new languages with minimum efforts and still to achieve an acceptable recognition rate. It is also important that the recognizer can easily be configured to different tasks, i.e. both the recognition vocabulary and grammar can be processed dynamically without any user involvement. This type of framework, illustrated in Figure 1, was proposed in [22].

Before converting the written vocabulary item into a sequence of spoken sounds, one first needs to perform text based language identification and choose accordingly the appropriate automatic text-to-phoneme conversion approach. These two automatic steps ensure that no user involvement is at all required when downloading the recognition vocabulary and grammar. Figure 1 also lists some basic techniques that can be applied to these tasks. Since automatic data-driven techniques are prone to errors, one should utilize application-specific knowledge, e.g. for language identification, if just possible. In wireless devices, the use of large pronunciation dictionaries is understandably only seldom feasible because of memory limitations.

The acoustic basis of the recognition system is a set of multilingual monophone HMMs that simultaneously support several languages and have been defined for example according to the International Phonetic Alphabet (IPA) [24]. This arrangement results in substantial savings in the system development costs. By sharing acoustic models across various languages it is possible to add new languages to the system without any additional training steps (provided that the multilingual phoneme set contains all necessary phonemes of a new language). Acoustic model adaptation techniques play an important part in multilingual ASR for compensating the mismatches between the training and testing conditions.

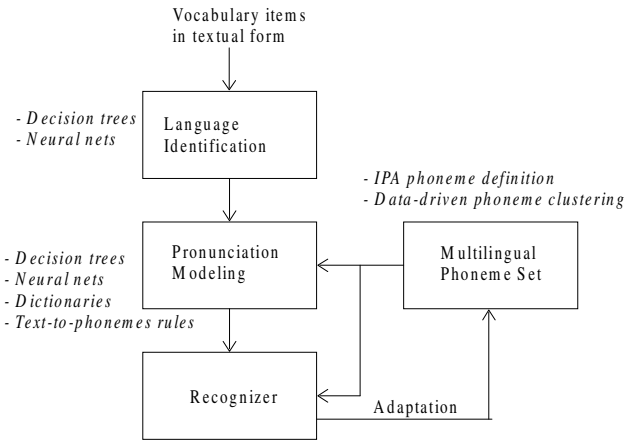

Figure 1: A multilingual ASR framework.

3.2.1. Multilingual ASR and Portable Terminals

In wireless devices, multilinguality aspects of ASR have a special importance. As the same systems, e.g. GSM phones, are sold over large geographic areas covering numerous different languages, it is preferred that there would be a "global" ASR engine integrated into terminals which would be capable of simultaneously supporting the required languages. The following characteristics of portable terminals justify the need to develop multilingual ASR technology:-

- Many languages spoken in the same area to where systems are offered
- Non-native users
- Non-native vocabulary items

By providing multilingual ASR systems, the development costs associated with localization and logistics can be minimized as there is no need to tailor several different language versions from the same recognizer. It is therefore preferred to include as many languages as possible in the same recognizer version.

Recognition of non-native speech is a new emerging challenge as speech recognition applications are becoming increasingly global. Voice browsing type of applications, e.g. the browsing of the Internet content by voice, contain lots of vocabulary items which do not necessarily match the user's native language. It is obvious that non-native pronunciation modeling techniques need to be developed as non-native speakers tend to insert or delete certain phonemes which are (not) spoken by native speakers. Recent studies have indeed shown the importance of non-native pronunciation modeling [3][18]. The acoustic realization of individual phonemes may also differ con-

siderably depending on the user's ability to speak the language. To compensate these variations, acoustic model adaptation techniques are again needed.

4. MAXIMIZATION OF RECOGNITION ACCURACY

A high recognition rate with all speakers across all possible operating conditions is of course a natural goal in all ASR applications. In real-world systems, and especially in embedded devices, it is nevertheless always essential to question each new algorithm in order to ensure that the performance gain obtained is worth the additional implementation costs. In this section, various ASR algorithms and methods running on wireless terminals are reviewed. This overview is not complete by any means, but the pros and cons of some of the most popular and recent techniques are here discussed. The reader is referred for example to [4][8] for more complete reviews of robust ASR techniques.

4.1. Noise Robustness

The portability of wireless devices sets very high technical challenges for the ASR technology. Since the devices can virtually be taken everywhere, noise robustness is a major issue to be considered in the algorithm design. This is particularly true when knowing that the recognition accuracy of the state-of-the-art speech recognizers tends to degrade even in the controlled noise conditions. In portable devices, SNR may vary between +30 and –10 dB and the background noise types can range from stationary to highly non-stationary noise signals. There is also no guarantee that two consecutive utterances would be spoken in the same, or even similar, noise conditions, as environment changes can occasionally occur abruptly. Due to limited battery life, it is also not technically feasible to attempt to continuously track noise characteristics using some on-line estimation techniques.

4.1.1. Noise Robust Feature Extraction

Ideally, all processing required to compensate for the harmful effects of ambient background noise would be done in the feature extraction stage. As of today, noise robust feature extraction typically includes methods which:-

- Aim at extracting parameters that are inherently robust to environmental variations, e.g. RASTA processing [6]. This has turned out to be difficult as there are today no such front-end algorithms which would produce truly noise immune features.
- Attempt to modify or normalize features to another representation where it is easier to cope with distortions, e.g. cepstral mean (and variance) normalization [14][21], or linear discriminant analysis [16].
- Attempt first to estimate the noise spectrum, and then to subtract this from the noisy input signal resulting in a clean speech estimate, e.g. all different variants of spectral subtraction.
- A combination of all above.

Low complexity is an undoubted advantage of all front-end based noise compensation methods. Computational complexity of almost all these techniques is so insignificant that there are usually no technical obstacles to include them in embedded devices. The major issue is in the performance column – there does not exist an advanced compensation algorithm which would maximize the recognition rate in all tasks across all possible noise conditions. While a high recognition rate can easily be achieved in one test environment, these positive results very seldom generalize to other operating conditions with completely different acoustics characteristics.

Many noise removal and spectral subtraction techniques require certain parameters to be estimated, e.g. a noise estimate, before the actual parameter compensation can be performed. This on-line estimation has turned out to be difficult in real-world ASR applications. First, there are no well performing solutions to carry out the noise estimation for non-stationary distortions which reduces the efficiency of the techniques. Second, accurate noise estimation is difficult as only very short segments of noise are usually available for estimation. The interaction time is usually short, e.g. 1-3 seconds in command & control type of ASR applications, and the system often needs to provide almost an immediate feedback to the user. It is thus obvious that accurate noise estimation is not feasible without introducing long processing delays and decreasing the usability of applications.

Despite the signigicant efforts put on noise robust feature extraction, fairly basic front-end algorithms are still today customary in real-world ASR applications. Cepstral coefficients, their first- and second-order time derivatives together with cepstral mean normalization have been to date, and will still likely be in the foreseeable future, the standard feature set on which the real-world ASR applications mainly rely. If the target environment is fixed and prior known, the use of highly optimized spectral subtraction algorithms has also found to be effective.

4.1.2. Noise Compensation in Acoustic Models

Since the extracted features are not resistant to background noise, numerous techniques have been proposed for carrying out noise compensation directly in acoustic models. The following basic families of techniques can be distinguished:-

- Multi-style (or multi-environment) trained acoustic models
- Parallel Model Combination (PMC) [1] type of techniques
- On-line adaptation of acoustic models

In practical systems, the PMC type of techniques suffer from the same limitation as many noise removal methods. The accurate estimation and modeling of non-stationary noises reduces the performance gains. Although some promising results have been obtained in a reasonably stationary car environment using PMC or Jacobian adaptation [15], the improvements are more moderate if more non-stationary noises are considered. Many model level noise compensation schemes, e.g. PMC, are also computationally costly which make them not very attractive solutions in many cases.

If the noise characteristics of an operating environment of an ASR system are known in advance, multi-style training is a good alternative to many more advanced noise compensation techniques. Multi-style training is simple and it performs relia-

bly as long as there is a reasonably good match between the training and testing conditions. In wireless devices, however, the assumption of a fixed and prior known operating condition is only seldom valid due to the high degree of portability of the devices. Multi-style training is therefore used mainly to optimize the system performance for specific acoustic conditions. Car environment is a typical example of an environment where the use of multi-style training has been found useful.

Acoustic model adaptation can also be considered as an on-line version of multi-style training. The HMM adaptation algorithms developed for speaker adaptation, such MAP and MLLR, can both be applied to environment adaptation as well. Environment adaptation has widely been reported to provide good performance improvement in various recognition tasks.

4.2. Acoustic Model Adaptation

In all speaker-independent speech recognition applications, not only in those running on embedded wireless systems, acoustic model adaptation techniques play an important role for maximizing the recognition accuracy. An exhaustive overview of the different model adaptation techniques can be found for example in [23]. As pointed out in various places in this paper, adaptation is required to carry out compensation along many directions in multilingual ASR applications.

First, speaker adaptation is needed to compensate the mismatch between the speaker-independent acoustic models and the speaker-specific pronunciation characteristics. Second, environment adaptation optimizes the acoustic models to match the operating environment(s) where the user is using the ASR application. Third, it is possible that the language spoken by the user is not seen during the initial training phase. Model adaptation is then required to adjust the parameters of multilingual HMMs for a new unseen language. Fourth, the initial acoustic model training has often been done using the publicly available databases (for minimizing the data collection burden), and therefore, the recognition task in testing does not necessarily match the data collection domain, e.g. continuous speech in training vs. isolated words in testing. Adaptation is thus also performing parameter optimization to an unseen recognition task domain. The role of acoustic model adaptation is crucial for maximizing the recognition performance for the given speaker, environment(s), language, and recognition task.

The major issue behind generally well performing adaptation techniques is that they require speech for carrying out parameter compensation. Particularly in environment adaptation, this reduces the performance gain obtained as the compensated models can only be used in the recognition of next utterance. It would therefore be useful if one could distinguish between environment and speaker adaptation. Quite recently, promising results have indeed been obtained by a joint use of Jacobian adaptation and MAP/MLLR techniques [12].

To make the adaptation process as transparent as possible to the user, unsupervised and on-line adaptation modes are preferred. Even though supervised adaptation is known to outperform its unsupervised counterpart, supervised adaptation should be used only if it can be realized in such a way that no user assistance is required. The same reasoning is also valid for off-line adaptation. It is very difficult to justify to the users the advantages of speaker-independent ASR technology if they still need to undergo lengthy enrollment sessions.

5. SUMMARY AND OUTLOOK

In this paper, the role of automatic speech recognition in portable wireless devices, particularly in mobile terminals, has been reviewed. Today, due to its low complexity and a high degree of robustness, the speaker-dependent technology is today, and also in the near future, the major ASR approach in mobile phones and wireless devices. Various technical issues required to make the transition from speaker-dependent to speaker-independent ASR were discussed. Clearly, there is a lot of potential behind speaker-independent ASR, but there are several technical obstacles to be solved, such as multilinguality, cost-efficient ASR language development, and noise robustness, just to name a few, until the wide-scale utilization of speaker-independent ASR is feasible.

Arguably, more advanced ASR applications and services will gradually appear for wireless devices. In terminals, the ease-of-use aspects continue to drive the introduction and development of new voice user interface applications. Since the characteristics of network and terminal ASR are very different to each other, it is obvious both network and terminal ASR applications will co-exist in the future. From the end-user's perspective, it is of course less important where the recognition engine is actually located, but the main issue is that the offered ASR services and applications provide clear tangible benefits for the users.

Instead of the rapid breakthrough, it is more likely that the appearance of more advanced ASR applications will occur only gradually, as there are no readily available solutions for various technical problems. Making significant progress in any of the key areas of ASR takes nowadays a lot of time. Furthermore, the first algorithmic solutions are usually nowhere near the capabilities of low cost wireless devices, and therefore, an additional R&D cycle is often needed to cut down the complexity of new technology while still maintaining the original performance gain. In addition to technology development, it is also essential to focus on analyzing the application and economical sides of ASR. It is important to understand what are the reasons behind the success or failure of certain ASR applications. Only the combined efforts of technology and application development ensure the realization of technically and economically viable ASR applications for wireless devices.

6. ACKNOWLEDGMENTS

The paper relies on research, observations, and general discussions that have jointly been carried out in Voice Interfaces group at Nokia Research Center. In addition to the whole group, the author particularly wishes to thank Juha Häkkinen and David Bye (both from Nokia Mobile Phones) as well as Ramalingam Hariharan, Pekka Kapanen and Juha Iso-Sipilä (all from Nokia Research Center) for their help and valuable comments for this paper.

7. REFERENCES

[1] M. J. F. Gales, S. J. Young, "Robust continuous speech recognition using parallel model combination", *IEEE Trans. on Speech and Audio Procssing*, Vol. 4, No. 5, pp.352-359, 1996.

[2] J.-L. Gauvain, C-H. Lee, "Maximum a posteriori estimation of multivariate Gaussian mixture observations of Markov chains", *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 2, pp. 291-298, 1994.

[3] S. Goronzy, R. Kompe, S. Rapp, "Generating non-native pronunciation variants for lexicon adaptation", *Proc. of ISCA Workshop on Adaptation Methods for Speech Recognition*, 2001.

[4] Y. Gong, "Speech recognition in noisy environments: A survey", *Speech Communication*, No. 16, pp. 261-291, 1995.

[5] P. Haavisto, "Audio-visual signal processing for mobile communications", *Proc. of EUSIPCO'98*, 1998.

[6] H. Hermansky, N. Morgan, "RASTA processing of speech", *IEEE Trans. on Speech and Audio Processing*, pp. 578-589, 1994.

[7] J. Iso-Sipilä, K. Laurila, R. Hariharan, O. Viikki, "Hands-free voice activation in noisy car environment", *Proc. of Eurospeech'99*, 1999.

[8] J.-C. Junqua, J.-P. Haton, *Robustness in Automatic Speech Recognition – Fundamentals and Applications*, Kluwer Academic Publishers, Boston, 1996.

[9] C. J., Leggetter,. P. C., Woodland, "Speaker adaptation of HMMs using linear regression", *Proc. of ICLSP'94*, 1994.

[10] M. Mohri, "Finite-state transducers in language and speech processing", *Computational Linguistics*, Vol. 23, No. 3, 1997.

[11] D. Pearce, "Enabling new speech driven services for mobile devices: An overview of the ETSI standards activities for distributed speech recognition", *Proc. of AVIOS'00*, 2000.

[12] L. Rigazio, D. Kryze, P. Nguyen, J.-C. Junqua, "Joint environment and speaker adaptation", *Proc. of ISCA Workshop on Adaptation Methods for Speech Recognition*, 2001.

[13] S. Riis, O. Viikki, "Low complexity speaker-independent command word recognition in car environments", *Proc. of ICASSP'00*, 2000.

[14] A. E. Rosenberg, C.-H. Lee, F. K. Soong, "Cepstral channel normalization techniques for HMM based speaker verification", *Proc. of ICSLP'94*, 1994.

[15] S. Sagayama, Y. Yamaguchi, S. Takahashi, "Jacobian adaptation of noisy speech models", *Proc. of IEEE ASRU Workshop'97*, 1997.

[16] O. Siohan, "On the robustness of linear discriminant analysis as a preprocessing step for noisy speech recognition", *Proc. of ICASSP'95*, 1995.

[17] V. Steinbiss, B.-H. Tran, H. Ney, "Improvements in beam search", *Proc. of ICSLP'94*, 1994.

[18] J. Tian, I. Kiss, O. Viikki, "Pronunciation and acoustic model adaptation for improving multilingual speech recognition", *Proc. of ISCA Workshop on Adaptation Methods for Speech Recognition*, 2001.

[19] M. Vasilache, "Speech recognition using HMMs with quantized parameters", *Proc. of ICSLP'00*, 2000.

[20] M. Vasilache, O. Viikki, "Speaker adaptation of quantized parameter HMMs ", *Proc. of Eurospeech'01*, 2001.

[21] O. Viikki, D. Bye, K. Laurila, "A recursive feature vector normalization approach for robust speech recognition in noise", *Proc. of ICASSP'98*, 1998.

[22] O. Viikki, I. Kiss, J. Tian, "Speaker- and language-independent speech recognition in mobile communication systems", *Proc. of ICASSP'01*, 2001.

[23] P. C. Woodland, "Speaker Adaptation: Techniques and Challenges", *Proc. of IEEE ASRU Workshop'99*, 1999.

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

[25] Nokia Annual Report 2000, available in http://www.nokia.com/investor/2000/4Q/