International Seminar On Non-Conventional Energy Sources for Sustainable Development of Rural Areas,
IJAERD- International Journal of Advance Engineering & Research Development
e-ISSN: 2348-4470, p-ISSN:2348-6406
Impact Factor: 4.72, DOI:10.21090/ijaerd

A Technological Survey Of Speech Recognition Techniques

[1]Bhupesh Deshmukh, [2]Sharon Chhatre

Abstract— *This paper presents an outline of various approaches for providing automatic speech recognition (ASR) technology to mobile users. Three principal system architectures in terms of employing wireless communication link are analyzed: Embedded Speech Recognition Systems, Network Speech Recognition (NSR) and Distributed Speech Recognition (DSR). Overview of the solutions that became standards by currently as well as some critical analysis of the most recent developments within the field of the speech recognition in mobile environments is given. Open problems, pros and cons of the various methodologies and techniques are highlighted. Special stress is created on the constraints and limitations the ASR applications encounter beneath the different architectures.*

Keywords*—automatic speech recognition, acoustic models, back-end, feature extractors, front-end*

I. INTRODUCTION

The past decade has witnessed an unprecedented developing of the telecommunication industry. Market researchers report that there are more than 1.6 billion mobile phone subscribers worldwide as of 2005 and this amount is expected to grow up to 2 billion by the end of 2006. The today's mobile technologies have far overcome person-to-person communication. The 2.5G networks supporting packet switched information exchange, like GPRS with realistic bit rates of 30-80 kbit/s, became an everyday matter. while the networks of the third generation, like UMTS or CDMA2000, ar already operated in 72 countries, have thirty two million users (middle 2005) and still unfold. These networks provide an effective data transfer rate upto 384 kbit/s.

At the same time the wireless local area networks (WLANs) based on the IEEE 802.11 specifications also known as Wi-Fi spots became widely available. This is a technology, which enables a person with a wireless-enabled computer or personal digital assistant (PDA) to communicate being within the coverage of an access point. With rates up to 11 Mbit/s Wi-Fi makes possible such applications as Voice over IP or video conferencing. Alongside with expansion of the network technologies, the client devices have been developing at the same speed. Nokia has forecasted that by the end of 2006 one sixth of all cellular phones will be UMTS supporting devices. Also PDAs are getting more and more popular. For example, according to Gartner's study, the overall market for PDAs has grown by 20.7% in the third quarter of 2005, compared to 2004.

Of course such a perfect infrastructure gave rise for the development of many new data services for the handheld devices. However, the user interface, which has definitely improved over the last years, still limits the usability of the mobile devices. The main interface problem of handheld gadgets is their miniature size. Typing on such tiny keyboards or pointing with stylus is very uncomfortable and error prone. Another issue is that PDA are often used when person is really "on the move". Operating in such conditions is either impeded or even prohibited, e.g. in the case of car driving. The natural way to solve this problem consists in using speech recognition technology. Speech input requires neither visual nor physical contact with devices. It can serve as an alternative interface to the regular one or be a complement modality speeding up the input process and increasing its fidelity and convenience. In the last decade a substantial effort has been invested in the automatic speech recognition techniques. As a result fast, robust and effective speech recognition systems have been developed [1–3]. The modern state-of-the-art ASR systems provide the performance quality (usually assessed by the recognition word error rate (WER) and by the ratio of the processing time to the utterance duration), which affords the comfortable use of ASR in real applications. However, the direct reproduction of the algorithms suitable for the desktop applications is either not possible or mostly leads to unacceptable low performance on the mobile devices. Due to the highly variable acoustic environment in the mobile domain and very limited resources available on the handheld terminals the implementation of such systems on the mobile device necessitates special arrangements [4].

In this paper we address the system optimization techniques, which enable the speech recognition technology on the portable computing devices. The remainder of the paper is organized as follows. The process of the speech recognition from the perspective of the handheld device is presented in section 2. Sections 3, 4 and 5 draw in detail three principal system architectures: *client-based*, *server-based* and the *client-server* ASR. Section 6 summarizes and closes the discussion.

Table 1: Recognition rates of the state-of-the-art desktop ASR systems [2]

.

Task	Words	WER %	Real Time (RT)	
			1-CPU	2-CPU
Connected Digits	11	0.55	0.07	0.05
Resource Manag.	1000	2.74	0.50	0.40
Wall Street Journal	5000	7.17	1.22	0.96

II. ARCHITECTURES OF ASR SYSTEMS FOR MOBILE DEVICES

In this section we briefly describe the functional blocks of the current state-of-the-art speech recognition engines with their impact on the design of the ASR systems for mobile applications.

A. The Mobile ASR Dilemma

The implementation of effective mobile ASR systems is challenged by many border conditions. In contrast to the generic ASR, the mobile recognition system therefore has to encounter the following aspects: limited available storage volume (language model and acoustic models to be shorten, which leads to performance degradation), tiny cache of 8-32KB and small and ”slow” RAM memory from 1MB up to 32MB (many signal processing algorithms are not allowed), low processor clock-frequency (enforces to use the suboptimal algorithms), no hardware-made floating point arithmetic, no access to the operating system for mobile phones (no low level code optimization possible), cheap microphones (often far away from the mouth - affects the performance substantially), highly challenging acoustic environment (PDA can be used everywhere: in the car, on the street, in large halls and small rooms; this introduces additive and convolutional distortions of the speech signal), no real PDA recorded speech corpora are currently available, high energy consumption during algorithms execution, and so forth. Finally, improvements which could be done in one functional block contradict with other parts of the system.

B. System Configurations for Mobile Speech Recognition

ASR systems can be decomposed into two parts: the acoustic *front-end*, where the process of the feature extraction takes place and the *back-end*, performing Viterbi search based on the acoustic and language models. Since most of the portable devices use a communication link, we can classify all the mobile ASR systems upon the location of the front-end and back-end. This allows us to distinguish three principal system structures:

1. **client-based** architecture or embedded ASR, where both front-end and back-end are implemented on the terminal;
2. **server-based** or network speech recognition (NSR), where speech is transmitted over the communication channel and the recognition is performed on the remote server;
3. **Client-server** ASR or distributed speech recognition (DSR), where the features are calculated on the terminal, whilst the classification is done on the server side.

Each approach has its individual shortcomings, which influence the overall performance. Therefore, the appropriate implementation depends on the application and the terminal properties. Small recognition tasks are generally recommended to reside on terminals, while the large vocabulary recognition systems take advantage of the server capacities. In following we analyze the problems associated with particular architecture in detail and examine the remedies against undesired effects.

.

III. EMBEDDED SPEECH RECOGNITION SYSTEMS

In the case of client-based or embedded ASR the entire process of speech recognition is performed on the terminal device (see Fig. 1). Embedded ASR is often the architecture of choice for PDAs. First, these client devices are somewhat more powerful compared to the mobile phones. Second, they are driven under well established operating system, like Windows Mobile 5.0, allowing easer software extension at different system levels. Third, PDAs have well known processor architectures, e.g. Intel X Scale, and there are some libraries and development kits optimized for a particular platform [6]. Finally, PDAs do not always have a wireless communication link available, so the remote speech recognition is rather unwelcome on PDA. The main advantage of the terminal based architecture relies in the fact that no communication between the server and the client is needed. Thus, the ASR system is always ready for use and does not rely on the quality of the data transmission. The most important issue for embedded ASR systems is the very limited system resources on the mobile device. For the embedded ASR design two implementation aspects need to be considered. These are the memory usage of the underlying algorithms and the execution speed [7]. To achieve reliable performance of embedded speech recognition system the modifications improving both criteria should be introduced in every functional block of ASR system.

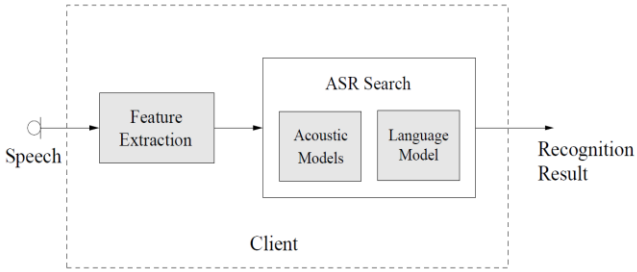

Figure 1: Client based ASR system - *Embedded Speech Recognition*

IV. NETWORK SPEECH RECOGNITION

Practically all complications caused by the resource limitations of the mobile devices can be avoided shifting both ASR frontend and back-end from the terminal to the remote server. Such a server-based ASR architecture is referred in the literature as network speech recognition. Unlike the embedded ASR, the NSR architecture can augment not only PDAs but also ”thin” terminals, e.g. cellular phones, with a very large vocabulary ASR. Another advantage of NSR relies in the fact that it can provide access to the recognizers based on the different grammars or even different languages. Besides, the content of the ASR vocabulary often may be confidential, thus prohibiting its local installation. Finally, the NSR allows a seem-less to the end-user upgrade and modification of the recognition engine. Characteristic drawback of the NSR architecture is the performance degradation of the recognizer caused by using low bit-rate codecs, which becomes more severe in presence of data transmission errors and background noise. To a certain extent the distortion introduced by source coding can be diluted if the recognizer is trained on the respectively corrupted speech. However, the tandeming of the different source coding schemes in addition to the different chanel noise levels spans a too large number of possible acoustic models. Better performance can be obtained if the recognition is performed based on the features derived from the parameteric representation of the encoded speech without the actual speech reconstruction. There are several successful implementations of such system intended for different codecs: ETSI GSM 06.10 [13], FS1015 and

FS1016 [14] and ITU-T G.723.1 [15]. Another important issue related to NSR design is an arrangement of the server side. In contrast to generic recognition systems, the NSR back-end should be able to serve effectively hundreds of clients simultaneously. In [16] Rose et al. suggest an event-driven, input-output non-blocking server framework, where the dispatcher, routing all the systems events, buffers the queries on the decoder proxy server, which redirects the requests to the one of free ASR decoder processes. Such NSR server framework composed of a single 1GHz proxy server and eight 1GHz decoder servers each running four decoder processes could serve up to 128 concurrent clients. [17] presents an alternative architecture, where the entire ASR system has been decomposed into 11 functional blocks. The components interconnected via DARPA Galaxy Hub can be accessed independently allowing a more efficient parallel use of the ASR system.

V. Distributed Speech Recognition

Distributed speech recognition represents the client-server architecture, where one part of ASR system, viz. primary feature extraction, resides on the client, whilst the computation of temporal derivatives and the ASR search are performed on the remote server. Even though both DSR and NSR make use of the serverbased back-end, there are substantial differences in these two schemes favoring DSR. First of all the speech codecs unlike the feature extraction algorithms are optimized to deliver the best perceptual quality and not for providing the lowest WER. Second, ASR does not need the high quality speech, but rather some set of characteristic parameters. Thus, it requires lower data rates - 4.8 kbit/s is a common rate for the features transmission. Third, since feature extraction is performed place on the client side, the higher sampling rates covering full bandwidth of the speech signal are possible. Finally, because in DSR we are not constrained to the error-mitigation algorithm of the speech codec, better error-handling methods in terms of WER can be developed. The studies within the distributed recognition framework target three aspects indicative for DSR:

1. the development of noise robust and computationally effective feature extraction algorithms;
2. the investigation of procedures for feature vectors quantization, permitting compression of the features without looses in recognition quality;
3. the elaboration of error mitigation methods.

VI. Conclusions

In this contribution we have analyzed three possible ASR architectures (embedded ASR, NSR and DSR) for providing the speech recognition technology to the portable end devices. The shortcomings associated with the particular design and possible solutions have been considered.

In our opinion the continuously increasing power of the mobile devices will give rise to the expansion of the embedded ASR systems. We suppose that within the near future the medium recognition tasks having 1000-2000 words, that represents an honest coverage of the sure application domain, will be with success running on the terminal devices, like PDAs or in-car embedded systems

In the light of the incremental affordability of high data-rate networks the remote speech recognition systems will also remain of interest for performing the very large vocabulary recognition and for accessing to corporate ASR system with confidential contents. Because of the superior performance of DSR in presence of the transmission errors and surrounding noise, with xAFE being selected by 3GPP for Speech Enabled Services, and with real time transmission protocol for AFE features standardized by IETF, we belive that NSR will be totally supplanted by the DSR architecture.

References

[1] B. Pellom and K. Hacioglu, "Sonic: The university of Colorado continuous speech recognition system," University of Colorado, Tech. Rep. TR-CSLR-2001-01, March 2001.

[2] W.Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," Sun Microsystems Laboratories, Tech. Rep. TR-2004-139, November 2004.

[3] J. Odell, D. Ollason, P. Woodland, S. Young, and J. Jansen, The HTK Book for HTK V2.0. Cambridge University Press, Cambridge, UK, 1995.

[4] R. C. Rose and S. Partharathy, "A tutorial on ASR for wireless mobile devices," in ICSLP, 2002.

[5] L. Rabiner and B.-H. Juang, Fundamentals of speech recognition. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[6] Intel, "Intel performance libraries," http://www.intel.com/cd/software/products/asmona/eng/perflib/index.htm, 2006.

[7] M. Novak, "Towards large vocabulary ASR on embedded platforms," in Proc. Interspeech 2004 ICSLP, 2004.

[8] T. W. K¨ohler, C. F¨ugen, S. St¨uker, and A. Waibel, "Rapid porting of ASR-systems to mobile devices," in Proc. Of the 9th European Conference on Speech Communication and Technology , 2005, pp. 233–236.

[9] A. Hagen, B. Pellom, and D. A. Connors, "Analysis and design of architecture systems for speech recognition on modern handheld-computing devices," in Proc. of the 11th International Symposium on Hardware/Software Codesign, October 2003.

[10] S. Ortmanns, T. Firzlaff, and H. Ney, "Fast likelihood computation methods for continuous mixture densities in large vocabulary speech recognition," in Proc. Eurospeech' 97, Rhodes, Greece, September 1997, pp. 139– 142.

[11] M. Vasilache, J. Iso-Sipil¨a, and O. Viikki, "On a practical design of a low complexity speech recognition engine," in Proc. ICASSP, vol. 5, 2004, pp. 113–116.

[12] M. Novak, R. Hampl, P. Krbec, V. Bergl, and J. Sedivy, "Two-pass search strategy for large list recognition on embedded speech recognition platforms," in Proc. ICASSP, vol. 1, 2003, pp. 200–203.

[13] J. M. Huerta, "Speech recognition in mobile environments," Ph.D. dissertation, Carnegie Mellon University, April 2000.
[14] B. Raj, J. Migdal, and R. Singh, "Distributed speech recognition with codec parameters," in Proc. ASRU'2001, December 2001.
[15] C. Pel´aez-Moreno, A. Gallardo-Antol´ın, and F. D´ıaz-de- Mar´ıa, "Recognizing voice over IP: A robust front-end for speech recognition on the world wide web," IEEE Trans. on Multimedia, vol. 3, no. 2, 2001.
[16] R. Rose, I. Arizmendi, and S. Parthasarathy, "An efficient framework for robust mobile speech recognition services," in Proc. ICASSP, vol. 1, 2003, pp. 316–319.
[17] K. Hacioglu and B. Pellom, "A distributed architecture for robust automatic speech recognition," in Proc. ICASSP, vol. 1, 2003, pp. 328–331.
[18] Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithm, ETSI Standard ES 202 050, October 2002.
[19] Recognition performance evaluations of codecs for Speech Enabled Services (SES), 3GPP TR 26.943, December 2004.
[20] T. Fingscheidt and P. Vary, "Softbit speech decoding: A new approach to error concealment," EEE Trans. On Speech and Audio Processing, vol. 9, no. 3, pp. 240–251, March 2001.
[21] V. Ion and R. Haeb-Umbach, "A unified probabilistic approach to error concealment for distributed speech recognition," in Proc. Interspeech 2005 ICSLP, 2005.
[22] A. James and B. Milner, "Soft Decoding of Temporal Derivatives for Robust Distributed Speech Recognition in Packet Loss," in Proc. ICASSP, vol. 1, 2005, pp. 345–348.
[23] K. K. Paliwal and S. So, "Scalable distributed speech recognition using multi-frame GMM-based block quantization," in Proc. Interspeech 2004 ICSLP, 2004