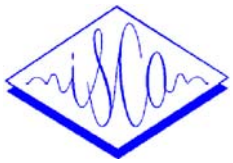
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

On the Integration of Speech Recognition into Personal Networks

Zheng-Hua Tan, Paul Dalsgaard and Børge Lindberg

SMC-Speech and Multimedia Communication, Department of Communication Technology
Aalborg University, Denmark
{zt, pd, bli}@kom.aau.dk

Abstract

Mobile communication presents a number of challenges to speech technology such as the limited resources available in the terminals in addition to the bandwidth constraints and the errors occurring in transmissions over mobile networks. These challenges need to be solved before automatic speech recognition (ASR) is ready for widespread use in the context of personal communication environments.

This paper gives an overview of the problems inherent in the recently developed network based ASR with an emphasis on the robustness issues that are highly influenced by network degradations. The paper further presents a number of transmission error protection and concealment schemes that are evaluated in a number of ASR experiments encompassing a range of typical real-environment transmission errors.

10.21437/Interspeech.2004-685

1. Introduction

Extensive R&D efforts in academia and industry presently address research issues aimed at communication in personal distributed environments – the so-called personal networks (PN) [1]. In such environments users interact with various companions, embedded, or invisible computers not only in their close vicinity but potentially anywhere. The vision of having PNs is that they comprise potentially all of a person's devices capable of network connection whether in her or his wireless vicinity, at home or in the office. The work towards enabling this vision transparently for users results in major extensions of the present personal area networking (PAN) and ambient intelligence (AmI) paradigms. At the heart of a PN is a core PAN, which is physically associated with the owner of the PN, as illustrated in Figure 1. Unlike PANs that have a limited geographically coverage, each PN has an unrestricted geographical span, which may incorporate devices into the personal environment regardless of their geographic location. A PN extends and complements the concept of pervasive computing.

In the PN environment there is a high demand for applications to include ASR as a key component of the user interface. However, the devices used by a PN owner are often hand-held devices with limited battery life, computing power and memory size for which reasons it is a challenge to implement any complicated ASR in the devices. In ASR associated with large databases the security and consistency considerations also make it hard to build ASR on the devices [2]. On the other hand, the 'always-on' facility of the PNs offers improved opportunities for running the ASR modules in a distributed architecture where only the front-end processing requires specific porting and implementation to the hand-held devices.

Successful functioning of ASR requires access to high or toll quality speech. In networked speech recognition, the performance of ASR will degrade seriously as the data may be infected with transmission errors and data packet loss. Lossy source coding causes degradation as well. Subsequently, the ASR modules require modifications to be able to function even for varying Quality-of-Service (QoS) of the transmission network.

This paper focuses on a number of problems that need be considered aiming at the deployment of robust ASR in PNs. Successful solutions are of central importance for the widespread use of ASR in PN based services that are expected to be part of the daily life for users of the future information society.

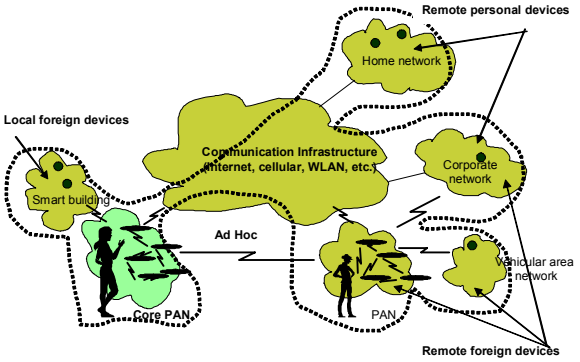


Figure 1*:* Illustration of the PN concept.

2. Speech recognition in personal networks

The integration of ASR into PNs can be implemented as either a purely terminal based or a network based architecture. Due to their different advantages both architectures will exist in the future. An overview and comparison of these architectures has been presented in [3]. This paper considers network based solutions only.

2.1. Network based speech recognition

The research within network based ASR is focused on three aspects, namely the architecture, source coding and channel robustness.

In a distributed framework for implementing ASR services on wireless mobile devices, efficiency is in focus to support a large number of mobile users connected via wireless network in [2]. In the DARPA Communicator program a more general distributed architecture for an ASR system has been developed incorporating a Hub and a variety of servers, and the ASR system is decomposed into a number of components allowing for flexible, efficient and effective interaction among them [4].

To enable efficient data transmission in distributed architectures source coding is applied to conduct data compression. Data representing speech is transmitted from the

input device to the server either as coded speech or as ASR features, resulting in two types of networked ASR namely server based ASR and distributed speech recognition (DSR).

In server based ASR the client samples the speech waveform and transmits the encoded speech only. The server re-synthesises the decoded data, conducts feature extraction and subsequently performs recognition. The quality of the re-synthesised speech is highly dependent on the speech coder. A low bit rate speech coder may cause significant degradations in recognition performance [5]. The effect of a variety of source coding schemes, e.g. voice over IP (VoIP) and GSM codecs, on ASR has been extensively investigated in [6], [7]. The feature set, however, may also be estimated directly from the bit-stream of the speech coder without synthesizing the coded speech [8].

In DSR speech features suitable for recognition are calculated and quantized in the client and transmitted to the server, where they are decoded, submitted to an appropriate error concealment (EC) algorithm and subsequently handled by the recogniser. This set-up provides a good trade-off between bit-rate and recognition accuracy [9].

To counteract the degradation in recognition performance due to a noisy channel, considerable research efforts have been conducted in exploring the potential of error protection and concealment techniques [9]-[16]. Channel robustness issues are further discussed in section 3.

Supported by the rapid growth in mobile communications, a number of algorithms that handle the combined effects of source coding, channel coding/decoding and EC techniques have been developed [9]-[11]. The ETSI published the first DSR standard with the aim of handling the degradations of ASR over mobile channels caused by both lossy speech coding and transmission errors [16]. The latest extensions to this are the advanced front-end aimed at noise-robust ASR and the extended front-end aimed at enabling improved tonal language recognition and server-side speech reconstruction by containing fundamental frequency information for the speech [17]. The DSR extended advanced front-end is under consideration by the 3rd Generation Partnership Project (3GPP) as a candidate for the speech enabled services (SES) codec, as an alternative to adaptive multi-rate (AMR) coding. Proposals have also been made to the Internet Engineering Task Force (IETF) to define Real-Time Protocol (RTP) payload formats for these DSR codecs [18].

2.2. The ETSI-DSR standard

The ETSI-DSR standard defines the feature-extraction front-end processing together with an encoding scheme [16]. The front-end produces a 14-element vector consisting of log energy (logE) in addition to 13 mel-frequency cepstral coefficients (MFCC) ranging from c_0 to c_{12} – computed every 10 ms. Each feature vector is compressed using split vector quantization (SVQ). The SVQ algorithm groups two features (either $\{c_i$ and c_{i+1}, i=1, 3...11$\}$ or $\{c_0$ and $logE\}$) into a feature-pair subvector resulting in seven subvectors in one vector. Each subvector is quantized using its own SVQ codebook. The size of each codebook is 64 (6 bits) for $\{c_i$ and $c_{i+1}\}$ and 256 (8 bits) for $\{c_0$ and $logE\}$, resulting in a total of 44 bits for each vector.

Two quantized frames – in this paper equivalent with vectors - are grouped together and protected by a 4-bit cyclic redundancy check (CRC) creating a 92-bit frame-pair. Twelve frame-pairs are combined and appended with overhead bits resulting in an 1152-bit multi-frame. Multi-frames are concatenated into a 4 800 bps bit-stream for transmission.

At the server two calculations determine whether or not a frame-pair is received with errors, namely a CRC test and a data consistency test. In the EC processing, a repetition scheme is applied to replace erroneous vectors.

The methods presented in this paper utilise the ETSI-DSR standard as a baseline.

3. Robustness against network degradations

Due to the alleviation of lossy source coding in the DSR framework, the remaining key robustness issue against network degradations is the presence of transmission errors. A number of solutions to overcome these problems have been developed on the basis of analyses of error characteristics.

3.1. Analysis of error characteristics

3.1.1. Burst-like vs. random

Errors occur not only due to channel noise but also due to a variety of transmission impairments. In real communication environments bit errors occur both with random and with burst-like distributions.

To investigate the distribution of errors, three GSM error patterns (EP) are analysed due to their realistic characteristic and to their common use in testing codecs. The EPs are segmented into frames each corresponding to 10 ms – the shift step of ASR feature extraction process. Each frame is then classified as error-free or erroneous depending on if there are any errors in the frame segment. The resulting distribution functions of erroneous frames as a function of length (number of consecutive erroneous frames) are shown in Figure 2.

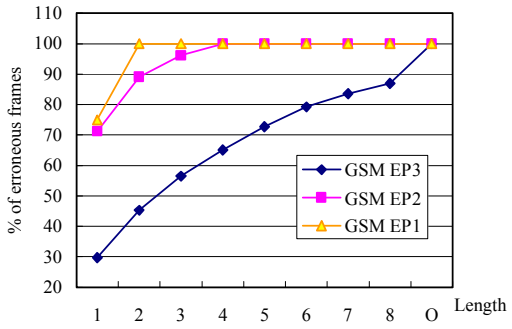


Figure 2: Distribution functions of erroneous frames by length. Length 'O' covers lengths larger than 8.

From Figure 2, it is noticed that the percents of erroneous frames of adding up all consecutive erroneous frames with a length from 1 to 3 (which are not considered as burst-like) are approximately 100%, 89% and 56% of the total erroneous frames for GSM EP1, EP2 and EP3, respectively. It indicates that both random and burst-like errors should be taken into account even though burst-like errors do more harm to ASR.

With the aim of counteracting especially burst-like errors the ETSI-DSR standard adopts a frame-pair scheme for error protection. However, the analysis above demonstrates that in reality random errors cannot be neglected. To take random errors into consideration as well, [13] proposed a one-frame error protection scheme where each frame is independently protected by a 4-bit CRC.

3.1.2. *Frame-pair vs. one-frame*

To compare the influence of applying the one-frame scheme as opposed to the frame-pair scheme, error rates of both schemes are calculated across a number of random bit error rates (BER) according to the following formula

$$ErrorRate = 1 - (1 - BER)^{bits} \quad (1)$$

where *bits* is the number of bits in the frame-pair or one-frame.

The results in Figure 3 show that the one-frame scheme significantly reduces the detected frame error rate (FER).

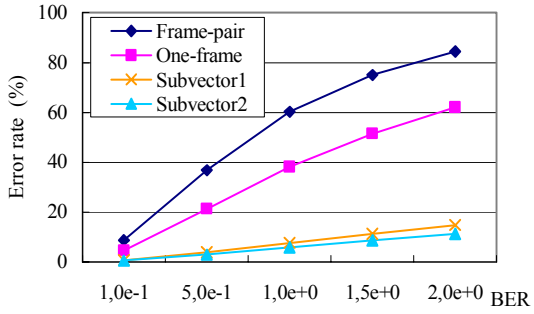


Figure 3: % Error rates of frame-pair, one-frame (vector) and subvectors vs. % BER.

3.1.3. *Vector based vs. subvector based*

In the ETSI-DSR standard, the EC is split into two parts where the first half of a series of erroneous frames is replaced with a copy of the last correct frame before the error and replacing the second half is replaced with a copy of the first correct frame following the error.

It is observed that the EC is conducted at the vector (or frame) level only. A vector is the unit selected for error detection, and if erroneous then followed by a substitution. This is the common characteristic of vector level EC algorithms no matter whether splicing, substitution, repetition or interpolation scheme is applied.

It is, however, highly likely that not all subvectors in an erroneous vector are corrupted by errors and the vector level EC strategy thus fails to exploit the error free fractions left within erroneous vectors.

To illustrate the potential of exploiting subvector information, vector and subvector error rates are calculated according to (1) as well where now *bits* is the number of bits in a vector or a subvector. The results are shown in Figure 3 where Subvector1 and Subvector2 are $[C_0, \log E]$ and $[c_i, c_{i+1}]$, i=1,3...11, respectively. It is noticed that the error rates of the subvectors are significantly lower.

3.2. Error-robust speech recognition

The results of the above analyses motivate a number of approaches to error-robust speech recognition.

3.2.1. *One-frame based error protection*

An error protection scheme based on the one-frame approach instead of the ETSI frame-pair based approach causes the overall probability of a frame being in error to be lower (at the cost of only a marginal increase from 4 800 bps to 5 000 bps in bit-rate) [13].

3.2.2. *Subvector based error concealment*

The exploitation of the potential error-free information embedded in each erroneous vector – rather than simply substituting them – leads to a subvector-level EC scheme in which each subvector is selected as the basis for supplementary error detection and mitigation.

Since there is no CRC coding applied at the subvector level, error detection at this level can only make use of a data consistency test.

Given that n denotes the frame number and $\boldsymbol{V}$ the feature vector, each vector is formatted as

$$\begin{aligned}\boldsymbol{V}^n &= [c_1^n, c_2^n \ldots c_{12}^n, c_0^n, \log E^n]^T \\ &= [[c_1^n, c_2^n] \ldots [c_{11}^n, c_{12}^n], [c_0^n, \log E^n]]^T \\ &= [[\boldsymbol{S}_0^n]^T, [\boldsymbol{S}_1^n]^T \ldots [\boldsymbol{S}_6^n]^T]^T \end{aligned} \quad (2)$$

where $\boldsymbol{S}_j^n$ (j=0,1…6) denotes the j'th subvector in frame n.

The consistency test is conducted across consecutive frame-pair vectors $[\boldsymbol{V}^n, \boldsymbol{V}^{n+1}]$ such that each subvector $\boldsymbol{S}_j^n$ from $\boldsymbol{V}^n$ is compared with its corresponding subvector $\boldsymbol{S}_j^{n+1}$ from $\boldsymbol{V}^{n+1}$. If any of the two decoded features in a feature-pair subvector does not possess a minimal continuity, the subvector is classified as inconsistent. Specifically both subvectors $\boldsymbol{S}_j^n$ and $\boldsymbol{S}_j^{n+1}$ in a frame-pair are classified as inconsistent if

$$(d(S_j^{n+1}(0) - S_j^n(0)) > T_j(0)) OR (d(S_j^{n+1}(1) - S_j^n(1)) > T_j(1)) \quad (3)$$

where $d(x,y)=|x-y|$ and $S_j^n(0)$ and $S_j^{n+1}(0)$ and $S_j^n(1)$ and $S_j^{n+1}(1)$ are the first and second element, respectively, in the feature-pair subvectors $\boldsymbol{S}_j^n$ and $\boldsymbol{S}_j^{n+1}$ as given in (2); otherwise, they are classified as consistent. The thresholds $T_j(0)$ and $T_j(1)$ are constants given on the basis of measuring the statistics of error free speech features. The threshold values given in the ETSI-DSR standard for data consistency test are used in the experiments for subvector based EC as given in Section 3.2.3.

The data consistency test generates a consistency matrix that discriminates between consistent and inconsistent subvectors. Only inconsistent subvectors are replaced by their nearest neighbouring consistent subvectors whereas the consistent subvectors are kept unchanged. Details are presented in [14].

3.2.3. *Experimental results*

To evaluate the performance of the methods presented above, two recognition experiments involving the Danish digits and city names were conducted. The experimental settings are as described in [13]. The baseline word error rate (WER) (no transmission errors) for the Danish digits and the city names are 0.2% and 20.7%, respectively.

The repetition scheme used by the ETSI-DSR standard was chosen to represent a set of alternative EC algorithms as comparison. This choice was taken due to the fact that the repetition scheme is a better safeguard to WER against transmission errors than methods like linear interpolation and splicing [12].

The GSM EPs were chosen to corrupt the speech feature stream. The GSM error patterns are EP1, EP2 and EP3 corresponding to carrier-to-interference ratios of 10 dB, 7 dB and 4 dB, respectively.

Figure 4 shows that the WERs for EP1 and EP2 for both tasks for all schemes are close to the baseline indicating that

the effect of EP1 and EP2 on recognition performance is insignificant. The degradation caused by EP3, however, is significant. But it is observed that the proposed one-frame scheme and subvector based EC technique significantly improve the performance.

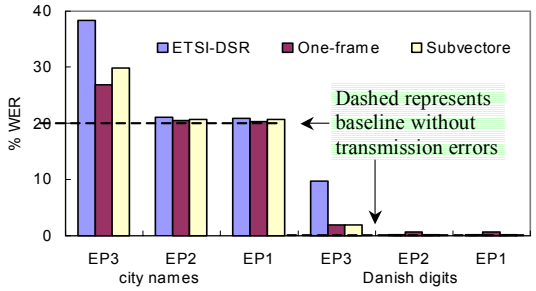


Figure 4: %WER for the Danish digits and city names for three schemes for the GSM EPs.

4. Adaptation to networks - towards QoS driven spoken language systems

In general each EC scheme is designed with the aim of maintaining maximal ASR accuracy. The front-end approaches, however, can only partly alleviate the impairment on speech features. In addition to the effort on the front-end solutions, the knowledge of network degradations can be further exploited and applied to the adaptation of the back-end recogniser.

One example of this kind is FER based out-of-vocabulary (OOV) detection [13] where it is observed that transmission errors influence the acoustic likelihood and thus affect the optimal threshold setting for discrimination between in-vocabulary words and OOV words. The CRC information in the channel error protection is exploited to calculate the current FER – a parameter of QoS – and a FER-dependent threshold that optimises the OOV detection can be estimated. This method has proved successful in maintaining a constant false rejection rate across a range of error rates.

To further extend this adaptation concept, QoS can be exploited to adapt e.g. the spoken language processing and dialogue management modules, with the aim of enabling graceful modifications to the behaviour of human computer interface (HCI). For example, the user can be requested to use a more restricted vocabulary and grammar or to switch to other modalities.

Furthermore, beyond individual best-effort research, cross layer design enables the interaction among different layers namely application, network and media access control (MAC) layer aiming at optimum QoS of the entire system [19].

5. Conclusions

This paper reviews the developments and trends of incorporating speech technology into a user-centric network architecture. Pervasive computing supported by PN offers improved opportunities for running ASR modules in a distributed architecture, enabling ASR deployment in a wide range of devices.

The research presented in this paper clearly shows the importance of applying a robust EC scheme to data transmitted across error-prone transmission channel. It is however pointed out that new research has to be initiated with the aim of introducing QoS-dependent modifications to existing ASR modules. The overall goal as seen from the users' perspective is to seamlessly offer robust and user-friendly HCI independent of which network is used.

6. Acknowledgements

This work is conducted in the context of the research project FACE (Future Adaptive Communication Environment), the consortium project CNTK (Centre for Network and Service Convergence) and the EU sixth framework project MAGNET (My personal Adaptive Global NET).

7. References

1. Niemegeers, I.G. and Heemstra de Groot, S.M., "From Personal Area Networks to Personal Networks: A User Oriented Approach", Personal Wireless Communications, Kluiwer Journal, May 2002.
2. Rose, R.C., Arizmendi, I. and Parthasarathy S., "An Efficient Framework for Robust Mobile Speech Recognition Services," Proc. ICASSP, Hong Kong, China, April 2003.
3. Viikki, O., "ASR in Portable Wireless Devices," Proc. ASRU, Madonna di Campiglio, Italy, December 2001.
4. Hacioglu, K. and Pellom, B., "A Distributed Architecture for Robust Automatic Speech Recognition," Proc. ICASSP, Hong Kong, China, April 2003.
5. Haavisto, P., "Speech Recognition for Mobile Communications," Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland, May 1999.
6. Kelleher, H., Pearce, D., Ealey, D. and Mauuary, L., "Speech recognition performance comparison between DSR and AMR transcoded speech," Proc. ICSLP, Denver, USA, Sept. 2002.
7. Mayorga, P., et al., "Audio packet loss over IP and speech recognition," Proc. ASRU, Virgin Islands, USA, 2003.
8. Kim, H.K. and Cox, R.V., "A bitstream-based front-end for wireless speech recognition on IS-136 communications system," IEEE Trans. SAP, vol. 9, pp. 558-568, July 2001.
9. Bernard, A. and Alwan, A., "Low-bitrate distributed speech recognition for packet-based and wireless communication", IEEE Trans. SAP, November 2002.
10. Boulis, C., Ostendorf, M., Riskin, E.A. and Otterson, S., "Gracefully degradation of speech recognition performance over packet-erasure networks," IEEE Trans. SAP, vol. 10, pp. 580-590, November 2002.
11. Potamianos, A. and Weerackody, V., "Soft-feature decoding for speech recognition over wireless channels," Proc. ICASSP, USA, May 2001.
12. Tan, Z.-H., Dalsgaard P. and Lindberg, B., "Partial splicing packet loss concealment for distributed speech recognition," *IEE Electronics Letters*, vol.39, no.22, pp. 1619-1620, October 2003.
13. ——, "OOV-detection and channel error protection for distributed speech recognition over wireless networks," Proc. ICASSP, Hong Kong, China, April 2003.
14. ——, "A subvector-based error concealment algorithm for speech recognition over mobile networks," Proc. ICASSP, Montreal, Quebec, Canada, May 2004.
15. James, A.B. and Milner, B.P., "An analysis of interleavers for robust speech recognition in burst-like packet loss," Proc. ICASSP, Montreal, Quebec, Canada, May 2004.
16. Distributed speech recognition; front-end feature extraction algorithm; compression algorithms, ETSI ES 201 108 v1.1.2 2000.
17. Ramabadran, T., et al., "The ETSI extended distributed speech recognition (DSR) standards: server-side speech reconstruction," Proc. ICASSP, Montreal, Quebec, Canada, May 2004.
18. "IETF AVT WG Internet-Draft RTP Payload Formats for ETSI ES 202 050, ES 202 211, and ES 202 212 Distributed Speech Recognition Encoding", IETF, February 2004.
19. Shakkottai, S, et al., "Cross-layer design for wireless networks," IEEE Communication Magazine, pp. 74-80, October 2003.