Comments from Stéphane H. Maes – smaes@us.ibm.com - +1-914-945-2908

Technical Specification Group Services and System Aspects **TSGS#13(01)0531**
Commented

Meeting #13, Beijing, China, 24-27 September 2001

Source:	**Siemens AG[1] - With Additional Comment**
Title:	**Comments on the time scale of proposed WI "Speech Enabled Services Based on Distributed Speech Recognition (DSR)" - Commented**
Document for:	**Discussion**
Agenda Item:	**7.1**

Abstract

The following comments attempt to clarify the status and the implications of the tasks that are involved in S1-010847 by commenting to SP-010488. This includes an enumeration of the tasks and specifications probably needed by a DSR framework (audio uplink and downlink streaming transport, session control, meta-information exchanges, uplink DSR optimised codecs). The activity is first and foremost the definition of a such DSR framework with a DSR protocol stack built on the same protocols as 3GPP and compatible with 3GPP rather than only a codec specification.

The experience of numbers of mobile equipment manufacturers and speech recognition vendors shows that a DSR framework is needed to deploy automated speech services as identified in S1-010847, S1-010846 and related DSR liaison documents and presentations (e.g. S1-010766, T2-010644). In the absence of such a framework, over 3G wireless networks, automated voice service performances may no be able to support levels of user experience needed for wide acceptance by users; despite the existing demand. In particular, the limitations of AMR/conventional codecs are discussed as part of the motivation for a DSR framework to support 3G automated speech services.

The mobile equipment manufacturers and speech recognition vendors that participate to 3GPP and support the DSR WI believe that a DSR framework can be specified and integrated within the 3GPP framework within the timescale of Release 5. This framework will fit 3GPP framework. For Release 5 the ETSI default DSR optimized codec can be used as default uplink codec while not limiting the framework to it. The DSR framework can be flexible to support future evolutions of codecs, DSR optimised codecs and associated standards. The comments discuss their support as negotiated codecs.

It is argued that the demand for automated speech services and the benefits of a DSR framework for operators and users are such that it should be made available as soon as possible within Release 5 rather than delayed to an unknown timescale of Release 6.

Commented Document

SA1 proposes a new work item called "Speech Enabled Services Based on Distributed Speech Recognition (DSR)" intended for Release 5. The goal of the WI is to define the necessary components for speech enabled services based on Speech Recognition, for example automatic speech access to information. Furthermore, it aims to identify the necessary changes and additions required in the current SA1 specifications. An invitation was sent to SA2, SA4, TSG-CN and T2 to participate in the work.
We appreciate the proposal towards the work on speech enabled services as part of the 3G specifications. With this contribution attention shall be drawn to the fact that as a Rel-5 work item two months (in case of December 2001 being Rel-5issuing date) or five months only (in case of March 2002 being Rel-5 issuing date) are left for the specification of an efficient speech enabled service solution.

[1] Dirk Didascalou, Siemens AG
Dirk.Didascalou@mch.siemens.de
Phone: +49 89 722 58574

Comments from Stéphane H. Maes – smaes@us.ibm.com - +1-914-945-2908

The ETSI STQ-Aurora DSR working group (Aurora) is developing a new DSR front-end that is planned to be decided for publication in Q2 2002.

There is no doubt that it is challenging to target DSR for Release 5.

However, there may be some confusion in terms of what is the status of the DSR activity at ETSI, the nature of the 3GPP DSR work item (S1-010847, 846) and the role of the "Aurora standards".

The 3GPP DSR work item and proposed specification is not a "codec specification and support" only WI. It is the specification of a framework that support efficient distribution of automated speech services within the network or on servers. To that effect a stack of protocols is needed to support:

- Streaming of the audio uplink and downlink back and forth between terminal and speech engines.
- Exchanges of:
  - session control protocols associated t the audio stream.
  - application specific and speech meta-information (e.g. keypad strings, display messages etc…)

Let us call such a protocol stack and framework a DSR framework. ETSI STQ Aurora DSR Application and Protocol working group has defined such a framework relying on IP protocols (based on IETF) expected to be directly compatible with the 3GPP stack. It builds on extensive experience and studies done by the ETSI Aurora members that include handset manufacturers and speech recognition vendors many of which are companies who are 3GPP members.

In principle such a framework does not impose the use of a particular codec (and it is not limited to 3G networks either). However, it is advantageous to use codec optimised for speech recognition. A discussion of such advantages is presented in [1] below.

ETSI (STQ-Aurora DSR) has published a first standard DSR optimised codec (ETSI ES 201 108). There are no other equivalent standardized DSR optimised codecs. This codec has been extensively tested for distributed speech services and proved to be a good compromised between bandwidth requirement, algorithm complexity and maintained speech recognition accuracy on a large spectrum of task perplexity.

It is possible to define other DSR optimised codecs better suited to particular engines, particular tasks or to better behave under certain acoustic conditions. Some speech vendors have such proprietary codecs. Also, ETSI Aurora is developing additional codec specifications (robust front-end, support for tonal languages and reconstruction) etc… There will always be in the near future a wide range of possible codecs and front-end algorithms that could be used and motivated for one reason or another. Nevertheless interoperability is best served by using a small number of standardised codecs.

The proposed DSR WI is to develop a DSR framework that includes:

- The DSR protocol stack as discussed above
- A default DSR codec for uplink exchanges as motivated in [1].
- Downlink/uplink streams for audio I/O.
- Support switches / negotiations of codecs (e.g. between a default DSR optimised codecs and other DSR codec variation like the robust front-end that ETSI will standardize in 2002 as discussed below). This of course would require that the terminal and server support such an alternate codec as discussed in [1].

Because of the amount of studies and expertise that is behind ETSI ES 201 108, and its adequate performances for automated speech services, and as in any case, this is only existing and agreed upon standardized optimised DSR codec, it should be considered as the default DSR codec or at least the starting point of any DSR optimised specification work.

Given the Rel-5 time frame, only the adoption of the new ETSI Aurora standard "as is" for a distributed speech recognition service seems to be possible. However, due to the planned publication of the ETSI Aurora Standard in 2002 this approach does neither guarantee timely adoption of the ETSI standard (It is a different codec than ETSI ES 201 108 that was published in February 2000. The codec targeted for 2002 is a robust front-end DSR codec. See the discussion above; the DSR WI does not target adoption of the "robust" or reconstruction aware DSR codecs. However, the resulting framework is expected to be able to support such codecs when available and supported by terminals and speech services via codec negotiation. Therefore, there are no issues of missing/non-finalized specification and the relationship between these different codecs is well understood.), nor the consistency with existing 3GPP specifications.

Based on the above discussion and framework, it should be clear that while requiring support for a DSR optimised codec, it does not lock the framework into limitations that may later be attributed to such a codec. On the other hand, it would allow Release 5 to support 3G automated speech services that will be acceptable to the users and should result into significant demand for such services.

Comments from Stéphane H. Maes – smaes@us.ibm.com - +1-914-945-2908

Most of the expected work, relies on supporting a DSR protocol stack within 3GPP. ETSI contribution is to be considered as an input and good starting point. The ETSI Aurora companies that are member of 3GPP believe that this can be done within the Release 5 time frame. Especially, if the targeted date for release 5 is 1Q 2002. Regarding the default optimised codec, there are not many other options than adopting ETSI ES 201 108 as is as default DSR optimised codec. There is no other candidate and there are no technical reasons to second-guess the technical work behind the specification of ETSI ES 201 108: it involved most of the speech vendors and handset manufacturers. It may be a significant burden on 3GPP to acquire the speech expertise required to specify another default DSR optimised codec and again, there are no technical reasons to expect that such a codec would be more or less optimised.

Speech recognition for speech enabled services might work also based on coded (e.g. AMR) speech with the full front-end and back-end part of the speech recogniser located in the network. (See discussion in [1]. From the speech recognition industry's experience, it will result into poor performances for medium to perplexity tasks. Good performances are not guaranteed even for low perplexity tasks.) In response to a UMTS Forum request from TSG-SA#12 (expressed in Tdoc SP-010294) SA4 states that (see Tdoc SA4 (01)0538):

> *"SA4 has not carried out any studies on the compatibility of speech codecs with voice recognition systems. There is no specific information available within SA4 which confirms or casts doubt on their suitability for use with network based voice recognition."*

i.e. a proof-of-concept still needs to be produced in order to ensure quality and usability of the speech enabled service. This is a statement about conventional codecs etc… based automated speech services. The statement above is that AMR has not been tested by SA4 for speech recognition. [1] explains that it is known that AMR codecs etc do not perform as well as what can be achieved with DSR optimised codecs and that the performance differences will make all the difference between acceptability/usability of the service and user frustration and irreversible disaffection for the services.

Really the SA#12 statement above confirms that, as elsewhere, automated speech services and speech recognition performances have never been a design point / requirement in the design of conventional codecs like AMR, G7.xx etc… As discussed in [1], the performances are rapidly poor (especially when compounded with the other challenges of mobile (m-commerce and others) voice services). If SA4 or another 3GPP working group wants to test this assertion and the limitations / issues raised in [1], it should be clear that numerous adaptation and training algorithms will have to be tested on numerous engine variations with numerous acoustic front-end variations. This is a huge task that has and is being carried by ETSI member companies and speech recognition companies and labs in general. It is quite understandable that SA4 decided against carrying such studies.

ETSI ES 201 108 is a working proven concept as is the proposed DSR framework.

As a consequence, we feel that the time scale of the WI should therefore be broadened, in order to ensure the specification of a consistent, correct and complete speech enabled service within 3GPP. For this purpose, it seems more feasible and relevant to open the respective WI for the time frame beyond Rel-5.

For future releases it may be appropriate to broaden the scope of the WI to address issues like:

- Speech services in general:
- Multi-modal interactions (see current proposal on T2 and SA-1 reflectors) [2].

However, this is clearly beyond the scope of what can be achieved within Release 5.

However, there is an immediate requirement (identified in the SA1 DSR liaison statements and WI draft See S1-010847 etc…) to provide the basic automated voice services to support that subscribers to the Distributed Speech Recognition Based Automated Voice Service are able to access information and conduct transactions by voice commands using distributed speech recognition (DSR). Example Automated voice services include:

- Basic voice services (Name dialing, Directory assistance, Interactive Voice Response system menus)
- Information retrieval (e.g., obtaining stock-quotes, checking local weather reports, flight schedules, movie/concert show times and locations)
- transactions (e.g., buying movie/concert tickets, stock trades, banking transactions)
- Personal Information Manager (PIM) functions (e.g., making/checking appointments, managing contacts list, address book, etc.)

Messaging (IM, universal messaging, etc…)

- Information capture (e.g. dictation of short memos)

DSR is today a viable way to deploy such automated speech service in a successful (usable and acceptable) way over 3G networks.

Comments from Stéphane H. Maes – smaes@us.ibm.com - +1-914-945-2908

In the absence of support by Release 5, such services on 3G networks will have very limited success without DSR support.

As described above, the companies involved believe that a flexible and extensible DSR framework can be integrated within 3GPP framework within the release timeframe. For Release 5 the ETSI default DSR optimized codec can be used as default uplink codec while not limiting the framework to it.

Eventually, the terminal support of DSR will most probably be optional within Release 5. Later releases it could be made it mandatory within certain conditions, if widely adopted.

From: Stephane Maes [smaes@US.IBM.COM]
Sent: Sunday, September 23, 2001 8:25 PM
To: 3GPP_TSG_SA@LIST.ETSI.FR
Subject: Comments to SP 010488 for SA#13 - Agenda item 7.1 - DSR

Importance: High

Dear colleagues,

Sorry for this late reaction to Dirk's document SP 010488. It was posted on a mail reflector that I do not closely monitor.

As it is directly relevant to the discussions that you are having this week at SA#13, I have attempted to clarify some of the issues related to the DSR WI that were raised in SP 010488.

As the comments are relatively long, I have also attached an abstract.

Feel free to contact me if you have any questions.

Sinrecely,

Stephane

Abstract

The following comments attempts to clarify the status and the implications of the tasks that are involved in S1-010847 by commenting to SP-010488. This includes an enumeration of the tasks and specifications probably needed by a DSR framework (audio uplink and downlink streaming transport, session control, meta-information exchanges, uplink DSR optimised codecs). The activity is first and foremost the definition of a such DSR framework with a DSR protocol stack built on the same protocols as 3GPP and compatible with 3GPP rather than only a codec specification.

The experience of numbers of mobile equipment manufacturers and speech recognition vendors shows that a DSR framework is needed to deploy automated speech services as identified in S1-010847, S1-010846 and related DSR liaison documents and presentations (e.g. S1-010766, T2-010644). In the absence of such a framework, over 3G wireless networks, automated voice service performances may no be able to support levels of user experience needed for wide acceptance by users; despite the existing demand. In particular, the limitations of AMR/conventional codecs are discussed as part of the motivation for a DSR framework to support 3G automated speech services.

The mobile equipment manufacturers and speech recognition vendors that participate to 3GPP and support the DSR WI believe that a DSR framework can be specified and integrated within the 3GPP framework within the timescale

Comments from Stéphane H. Maes – smaes@us.ibm.com - +1-914-945-2908

of Release 5. For Release 5 the ETSI default DSR optimized codec can be used as default uplink codec while not limiting the framework to it. The DSR framework can be flexible to support future evolutions of codecs, DSR optimised codecs and associated standards. The comments discuss their support as negotiated codecs.

It is argued that the demand for automated speech services and the benefits of a DSR framework for operators and users are such that it should be made available as soon as possible within Release 5 rather than delaying it to an unknown timescale of Release 6.

(See attached file: Comments_SP-010488_sm_9_23_01.zip)

___

Dr. Stéphane H. Maes,
Manager Mobile Speech Solutions and Conversational Multi-modal AdTech
IBM T.J. Watson Research Center,
P.O. Box 218, Yorktown Heights, NY 10598, USA.
Ph: (914)-945-2908; TL: 862; Fax: (914)-945-4490

>-----Original Message-----
>From: Didascalou Dirk ICM MP TI 5
>[mailto:Dirk.Didascalou@MCH.SIEMENS.DE]
>Sent: Friday, September 21, 2001 10:39 AM
>To: 3GPP_TSG_SA@LIST.ETSI.FR
>Subject: SP 010488: "Comments on the time scale of proposed WI "Speech
>Ena bled Services Based on Distributed Speech Recognition (DSR)",
>Document fo r SA#13
>
>Dear colleagues,
>
>attached please find the document:
>
> Title: Comments on the time scale of proposed WI "Speech Enabled
Services
>Based on Distributed Speech Recognition (DSR)"
> Source: Siemens
>
>the document is for discussion and should be dealt with under Agenda item
>7.1.

>Regards
>
>Dirk Didascalou
>_____
>Dr. Dirk Didascalou
>SIEMENS AG
>ICM MP TI5
>P.O.Box 801707 Tel: +49-89-722 58574
>D-81617 Munich Fax: +49-89-722 37078
>Germany Mobile: +49-160-4715 418
>mailto:Dirk.Didascalou@mch.siemens.de
>
> <<SP-010488.zip>>

---

[1] Why DSR Optimized Codecs?

It is important to carefully consider why accuracy of speech recognition is so relevant and by what it is affected. Hopefully this may put the issue in perspective and shift the focus from only transport error rates to also acoustic distortions introduced at encoding, etc... Acoustic front-end encoding is especially important to address the latter.

Comments from Stéphane H. Maes – smaes@us.ibm.com - +1-914-945-2908

First of all, with current speech technologies, accuracy depends on the complexity of the task, background environment and speaker characteristics. Errors are common and typically frustrating for users that have little tolerance for "machine mistakes that does not understand them". Acceptability of a service always depends on getting the highest possible accuracy and designing the application with the knowledge that errors will be made. This helps to present the user with the best user experience, even when he or she has to correct speech recognition mistakes.

Mobile environments are challenging because:

1) The acoustic environment / background noise can be very wild, uncontrolled, unpredicted or unknown.
2) Channel transmission errors especially in weak signal conditions can significantly degrade recognition performance using convention voice channels.
3) The user can be distracted and far from being a focused, calm and relaxed speaker
4) The tasks can be quite challenging and vary a lot in perplexity. Therefore, it is critical to provide the highest achievable accuracy for automated speech service to be widely accepted.

From the discussion within the different 3GPP groups, it is apparent that packet error used in the DSR framework will not significantly degrade the performances. This is good news. But there is another key factor to take into consideration. Speech recognition accuracy is directly affected by the nature and quantity of relevant spectral/acoustic information that is passed to the engine. Conventional codecs distort the waveform and spectrum in ways that are perceptually acceptable, but may often be dramatic at the level of speech recognition acoustic features. In other words, MEL Cepstra (the acoustic feature to extract from the waveform and pass as input to the speech recognition engine - virtually any state of the art speech recognition engine uses MEL cepstra or variations) are badly enough distorted to noticeably and irreversibly degrade the performance of the recognizer. And such degradation increases relatively when going for low vocabulary / low perplexity tasks to high perplexity tasks.

Typically with conventional perceptual codecs (e.g. AMR, GSM, G711.xx etc...) , it is possible to improve accuracy by training / adapting the acoustic model to the codec. However, our experience is that this never matches the performances that can be achieved with the non-distorted MEL cepstra. In addition, it may result into nightmarish bookkeeping and data file requirements for the service provider). Again, while achievable on low perplexity tasks this may become a showstopper on high perplexity tasks.

The compound effects of acoustic distortions, codec distortion and whatever small transport errors, has to be addressed. 3G networks etc. by enabling voice as data allows mechanism to reduce these effects:

- A codec / encoding scheme designed to minimize the ceptrsal distortions address the information loss issues and can still work at pretty low bit rates
- An appropriate encoding / error correction mechanism can handle whatever low packet losses are introduced (even if not important on 3G networks)
- Acoustic environment can be handled two ways:
    - By appropriate processing within the speech engine (trained acoustic model, compensation techniques etc...)
    - By appropriate front-end processing prior to generating the acoustic features. This can be addressed by 'negotiating" a particular DSR codec (or changing some of its settings) – provided that a terminal supports such an alternate codec.
        - The framework above will allow support of other DSR optimized codecs not necessarily “standardized by 3GPP – at least within release 5). This would be at the discretion of the terminal manufacturers and agreements between automated speech service providers, carriers and terminal manufacturers.

[2] See T2-010652 and T2-010705 presented at T2 and posted for e-mail discussions to SA-1 http://list.etsi.fr/scripts/wa.exe?A2=ind0109&L=3gpp_tsg_sa_wg1&F=&S=&P=786 .