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

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

STANDARDISED SPEECH INTERFACES - KEY FOR OBJECTIVE EVALUATION OF RECOGNITION ACCURACY

Christel Müller and Karsten Schröder
Deutsche Telekom Berkom GmbH, Research Section Speech Systems, Computer Telephony Systems
10589 Berlin, Goslarer Ufer 35, Germany
c.mueller@berkom.de, k.schroeder@berkom.de

10.21437/Eurospeech.1999-227

ABSTRACT

Today existing speech recognisers differ not only in their technology from single word to natural language understanding, from hardware based and software-only solutions but also in the recognition accuracy under real-time conditions in the whole variety of environments and networks. Standardised APIs for a defined state diagram of recognition process are the key requirement for integrating different technologies and their evaluation.
First time evaluation tests based on standardised interfaces and optimised vocabularies were carried out in an open system environment at Deutsche Telekom's Research Lab.
The approach of a unique lexicon development system enhances the whole process of vocabulary generation.
Optimising the phoneme transcription is one of the main future issues at the lab.

1. INTRODUCTION

Existing application development tools differ significantly in the integration of ASR recognition technologies, in the handling of vocabularies and grammars for the same dialogue state.
The main conditions for getting independent and comparable results of recognition accuracy are on a unique set of speech data one hand and a unique set of ASR functions at the resource layer on the other hand. Furthermore the development of vocabulary, grammar rules and parameter settings have to correspond for one selected test application.
Therefore a recogniser evaluation system configuration for unified interfaces of different ASR technologies was set up and several tests for key word spotter were carried out at the Deutsche Telekom research lab.
Additionally the differences in phonetic transcriptions of vocabularies led to first conclusions for an optimised deployment of ASR technology of German Language.
Finally the characteristics of speech data were described for this research approach of testing recognition accuracy under different conditions in telecommunication networks.
It was important to keep the balance between specific features of speech interfaces and restrictions of voice processing technology for the test application, since both factors influence and enhance each other in the process of technological advance.

2. STANDARDISED SPEECH INTERFACES

The ASR resource management layer has the task to implement unified ASR interfaces and accordingly to achieve vendor independence. This includes the recognition functions as well as the vocabulary, grammar and parameter settings for hardware and software based solutions.
The standardisation of ASR testing was carried out according to the S.100 and S.300 Interfaces, which were created by the members of ASR Task Group of Enterprise Computer Telephony Forum. These interfaces correspond to the existing SAPI and JSRAPI standards in terms of main functionality.

2.1. Subset of Functionality for ASR Testing

The independence of basic recognition functions and a selected set of parameters [1] are realised for tree selected technologies:

- HMM-Keyword Spotter RA, hardware based, multilingual
- HMM-Keyword Spotter RB, hardware based, multilingual
- HMM-Keyword Spotter RC, software based, multilingual. The handling parameters not only includes the initial settings but also the modification of their returned values and a repetitive setting:

„Recognize(Group,nNumResults,RTC,
OptArgs,TranInfo,Mode)"
„RetrieveRecognition(Group,ResulType,
Results,TranInfo,Mode)".

The 'Talk over' functionality was not taken into account for the different experiments.

2.2. Implementation into an open System Architecture

ASR evaluation in an open environment is a completely new approach to obtain true evaluation results under real conditions. An open system is characterised by:
variable choice of ASR technology provider,
multiple applications running at this environment,
standardised interfaces as well as administration and minimal effort of resource implementation.
Such an open system architecture consists of :

- User Interface Layer
- Network Layer
- Resource Layer
- Resource Management Layer and
- Application Layer,

where the resource layer contains the different ASR technologies and ensures the communication via a standardised speech interface.

3. DESIGN OF THE ASR EVALUATION SYSTEM

The approach covered ‘black box’ and ‘glass box’ aspects as well. Thus, the term evaluation system was used according to the definitions of assessment and evaluation given in [2].

The fully automated system for the speech recogniser assessment was built of the following components
(s. *figure 1*):

- Test application A, B
- Service Creation Environment with a graphical user interface,
- Computer telephony middleware for resource management of ASR and telephony and media processing,
- S.300 drivers for ASR and media control
- media control unit
- ASR software and hardware
- Primary Rate Interface (PRI) running in network mode
- Primary Rate Interface running in terminal mode
- PRI back-to-back cable.

Using the back-to-back dial mode via the digital telephone interfaces the timeslots on the SCBus for the speech recognition and for the playback of the speech samples were connected.
A special feature of the CT-software allows to start a play action and a recognition action in parallel within a single application. Thus, there was no need to syncronize two sites as shown in [3] and [4].
The speech coding format was PCM A-Law, 64 kBit/s. No conversions took place on the signal way in the system.
All recognisers got the identical digital speech input over the same interface (SCBus).

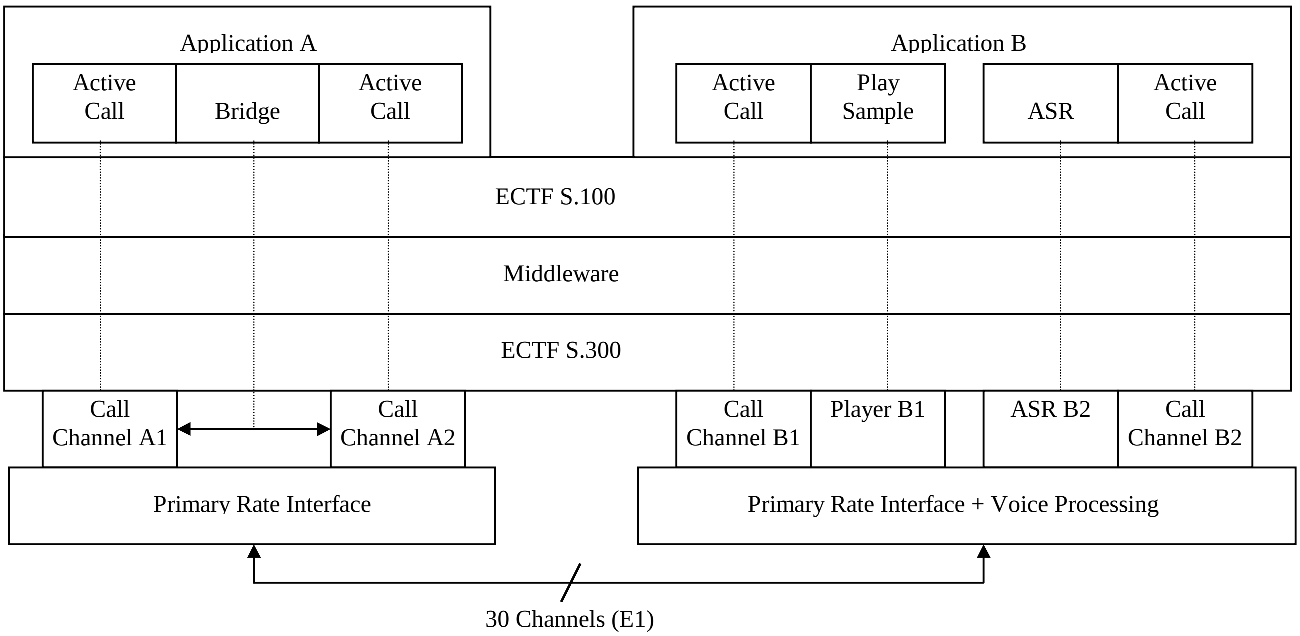

Figure 1: System Architecture of ASR Evaluation

4. EXAMPLE OF A COMPARATIVE ASR EVALUATION TASK

All three technologies were tested under the same environment with the same application.
Two test dialogues were started. After receiving and answering the call of dialogue 1 the second application played the test set of words and recognized the speech patterns in parallel.

4.1. Characteristics of Speech Data for objective ASR evaluation

Deutsche Telekom created a pool of speakers which covered all German dialects. The following information were available:

- gender,
- age,
- primary and secondary language
- primary and secondary dialect.

Speech samples read from word lists were collected for large telecommunication applications over the phone.
The features of every session were registered concerning:

a) the telephones characteristics:
- analogue standard and wireless,
- ISDN standard and wireless / DECT,
- analogue mobile / TACS,
- digital mobile / GSM,
- public telephone,

b) the exact name of the telephone and speakers' environments:
- car,
- home,
- office,
- landscape,
- street,
- telephone booth,
- public building and other.

Every speech sample was recorded into one separate speech file, 64 kbit/s PCM format, A-law without any conversions directly from the ISDN. After the validation process these recordings included defined periods of leading and trailing silence (approximately 200 ms). Certain subjective criteria concerning the quality were registered, e.g. kind of noise, recording level and intelligibility.

All features regarding the speakers, telephones, environments and subjective evaluations on the quality of the speech samples enable in-depth analyses of the recogniser's accuracy.

4.2. Design and Transcription of ASR Vocabulary

Design of ASR vocabularies in telecommunication applications depends on the possibilities of implementing synonyms for mission critical words or expressions.
The constant part normally consist of digits and names while the variety of word design only exists for dialogue control and interaction.
For the creation of a unique testing vocabulary a set of critical words was selected (s. *table 1*).
The phonetic transcriptions were processed automatically by the lexicon software. Differences between the transcriptions are shown in bold characters.
Characters representing silence or tone group boundaries were left out at the beginning and the end of the transcriptions strings.
If the specific lexicon software used product specific characters or IPA characters for the transcription, a translation to SAMPA was done according to [4].

Differences within the transcription alphabets existed as follows:

- Syllable boundary **$** (RC),
- Separator **-** (RC),
- Tone group boundary **|** (RB),
- Primary stress **“** (RB, RC)
- glottal stop **?** (RB, RC)
- near-open central unrounded /r/ **6** (RB)
- Syllabic /n/ **=n** (RB)
- trill/fricative **R** (RC, RB).

4.3. Accuracy Results

The tests of accuracy were carried out by 4050 German speech samples without using out-of-vocabulary data.
Automatically generated recogniser results are shown in *table 1*.
According to the provided tests of three different ASR technologies we concentrated on the differences in transcription of their associated vocabularies.
It turned out that the occurrence of phonetic specifics did not correlate to the deviations in recognition accuracy results.

<table>
<tr><th></th><th colspan="2">Recogniser A</th><th colspan="2">Recogniser B</th><th colspan="2">Recogniser C</th></tr>
<tr><th>Word</th><th>Transcription</th><th>WAcc</th><th>Transcription</th><th>WAcc</th><th>Transcription</th><th>WAcc</th></tr>
<tr><td>Amerika</td><td>ame:ri:ka:</td><td>99,33%</td><td>?am“e:|ri:|ka:</td><td>97,33%</td><td>a$“me:$rI$:ka</td><td>98,67%</td></tr>
<tr><td>Australien</td><td>aUstra:li:@n</td><td>100,00%</td><td>?aU|stRa:|li@n</td><td>100,00%</td><td>aUs$“tra:$lj@n</td><td>100,00%</td></tr>
<tr><td>Belgien</td><td>bElgi:@n</td><td>99,33%</td><td>b“El|gi:|@n</td><td>96,00%</td><td>“bEl$gi:$@n</td><td>97,33%</td></tr>
<tr><td>Brasilien</td><td>brazi:li:@n</td><td>100,00%</td><td>bRa|zi:|li@n</td><td>98,67%</td><td>bra$“zi:l$j@n</td><td>98,00%</td></tr>
<tr><td>Dänemark</td><td>dE:n@mark</td><td>100,00%</td><td>d“E:|n@|mark</td><td>96,00%</td><td>“de:$n@$maRk</td><td>100,00%</td></tr>
<tr><td>Deutschland</td><td>dOYtSlant</td><td>100,00%</td><td>d“OYtS|lant</td><td>100,00%</td><td>“dOYtS$lant</td><td>100,00%</td></tr>
<tr><td>England</td><td>ENlant</td><td>85,33%</td><td>?“EN|lant</td><td>83,33%</td><td>“EN$lant</td><td>93,33%</td></tr>
<tr><td>Finnland</td><td>fInlant</td><td>100,00%</td><td>f“In|lant</td><td>95,33%</td><td>“fIn$lant</td><td>91,33%</td></tr>
<tr><td>Frankreich</td><td>fraNkraIC</td><td>100,00%</td><td>fr“aNk|raIC</td><td>98,67%</td><td>“fraNk$raIC</td><td>93,33%</td></tr>
<tr><td>Griechenland</td><td>gri:C@nlant</td><td>100,00%</td><td>gr“i:|C@n|lant</td><td>96,00%</td><td>“gri:$C@n$lant</td><td>100,00%</td></tr>
<tr><td>Großbritannien</td><td>grOsbri:tani:@n</td><td>100,00%</td><td>gRo:s|bRi|ta:|ni@n</td><td>98,67%</td><td>gro:sbrI“tan$j@n</td><td>99,33%</td></tr>
<tr><td>Holland</td><td>hOlant</td><td>100,00%</td><td>h“O|lant</td><td>98,00%</td><td>“hOl$ant</td><td>99,33%</td></tr>
<tr><td>Irland</td><td>Irlant</td><td>95,33%</td><td>?I6|lant</td><td>64,67%</td><td>“IR$lant</td><td>72,67%</td></tr>
<tr><td>Italien</td><td>i:ta:li:@n</td><td>99,33%</td><td>?i:|t“a:|li:|@n</td><td>98,67%</td><td>i:$“ta:$lj@n</td><td>98,67%</td></tr>
<tr><td>Japan</td><td>ja:pan</td><td>100,00%</td><td>j“a:|pan</td><td>100,00%</td><td>“ja:$pa:n</td><td>98,67%</td></tr>
<tr><td>Kanada</td><td>kanada:</td><td>100,00%</td><td>k“a|na|da:</td><td>98,67%</td><td>“kanada</td><td>99,33%</td></tr>
<tr><td>Luxemburg</td><td>lUksEmbUrk</td><td>100,00%</td><td>lU|ks@m|bU6k</td><td>99,33%</td><td>“lUk$s@m$bURk</td><td>100,00%</td></tr>
<tr><td>Niederlande</td><td>ni:d@rland@</td><td>100,00%</td><td>ni:|d@6|lan|d@</td><td>99,33%</td><td>“ni:$d@R$lan$d@</td><td>100,00%</td></tr>
<tr><td>Norwegen</td><td>nOrve:g@n</td><td>100,00%</td><td>n“Or|ve:|g=n</td><td>96,67%</td><td>“nOR$ve:$g@n</td><td>99,33%</td></tr>
<tr><td>Österreich</td><td>2:st@raIC</td><td>96,67%</td><td>?“2:s|t@|raIC</td><td>96,67%</td><td>“2:s$t@$raIC</td><td>99,33%</td></tr>
<tr><td>Portugal</td><td>pOrtu:gal</td><td>98,67%</td><td>p“Or|tu:|gal</td><td>95,33%</td><td>“pOR$tu:$gal</td><td>99,33%</td></tr>
<tr><td>Rußland</td><td>rUslant</td><td>99,33%</td><td>r“Us|lant</td><td>92,00%</td><td>“rUs$lant</td><td>89,33%</td></tr>
<tr><td>Schweden</td><td>Sve:d@n</td><td>100,00%</td><td>Sv“e:|d=n</td><td>99,33%</td><td>“Sve:$d@n</td><td>100,00%</td></tr>
<tr><td>Schweiz</td><td>SvaIts</td><td>98,67%</td><td>Sv“aIts</td><td>98,67%</td><td>“SvaIts</td><td>98,67 %</td></tr>
<tr><td>Spanien</td><td>Spa:ni:@n</td><td>100,00%</td><td>Sp“a:|ni:|@n</td><td>100,00%</td><td>“Spa:n$j@n</td><td>100,00%</td></tr>
<tr><td>USA</td><td>u:Esa:</td><td>98,67%</td><td>?“u:|?“Es|?“a:</td><td>94,00%</td><td>u:Es?a:</td><td>100,00%</td></tr>
<tr><td>Vereinigte Staaten</td><td>fEraInIkt@sta:t@n</td><td>99,33%</td><td>fE6|?aI|nIC|t@|St“a:|t=n</td><td>97,33%</td><td>fEr$“aI$nIC$t@-“Sta:t@n</td><td>99,33%</td></tr>
</table>

Table 1: Phonetic Transcription of critical Words for a special Test Set vs. Word Accuracy

5. CONCLUSIONS

The research work at Deutsche Telekom Speech Systems, Computer Telephony Systems section, showed that the unified ASR interface with standardised functionality remains the only way of getting reasonable results in ASR assessment and evaluation of different technologies.

Due to the fact that no direct relation between the phonetic specifics and the deviations in accuracy could be found further experiments based on vocabularies transcribed in a unique manner will be provided. If for each recogniser these new test results will not differ significantly from the accuracy shown in *table 1* then there is no need to insist on technology provider dependent transcription schemes.

Deploying different recognition technologies for one application would only require one single vocabulary creation environment.

Today's ASR real-time applications only meet the user acceptance, if the described well defined vocabularies for voice user interfaces are integrated.

6. REFERENCES

[1] Enterprise Computer Telephony Forum (1996), S.100 Revision 1.0, Media Services "C" Language, Application Programming Interfaces.

[2] David S. Pallett, Adrian Fourcin (1995), Speech Input: Assessment and Evaluation. In A. Cole (editor), Survey of the State of the Art in Human Language Technology Ronald. http://cslu.cse.ogi.edu/HLTsurvey/.

[3] A. Fourcin, G. Harland, W. Barry & V. Hazan, (1989), Speech input and output assessment. Multilingual methods and standards. Ellis Horwood Ltd., Chichester.

[4] Dafydd Gibbon, Roger Moore and Richard Winski (eds.) (1997), Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, Berlin, New York.