NICE project (IST-2001-35293)

Natural Interactive Communication for Edutainment

Speech recognition and synthesis, natural language understanding robustness

NICE NISLab Extra Deliverable

28 September 2004

Author

Niels Ole Bernsen

1: NISLab, Odense, Denmark

Project ref. no.	IST-2001-35293
Project acronym	NICE
Deliverable status	Confidential
Contractual date of delivery	No date specified in contract 1 October 2004 requested at the NICE 2nd year review
Actual date of delivery	28 September 2004
Deliverable number	N/A
Deliverable title	Speech recognition and synthesis, natural language understanding robustness NISLab Extra Deliverable
Nature	Report
Status & version	Final
Number of pages	15
WP contributing to the deliverable	3
WP / Task responsible	NISLab is the sole responsible for the present deliverable
Editor	N/A
Author(s)	Niels Ole Bernsen
EC Project Officer	Mats Ljungqvist
Keywords	Speech recognition and synthesis, natural language understanding
Abstract (for dissemination)	N/A Please do not disseminate!

Table of Contents

1 Introduction

The present deliverable is an addition to the deliverables planned in the NICE project Work Plan as part of the Technical Annex to the NICE contract. The present deliverable only addresses the work done at NISLab in the NICE project. The deliverable was requested in the review report received by mid-June 2004 following the 2nd project review on 14 May 2004. The formulation of the request was the following:

> *3. Integrate State-of-the-Art speech synthesis and recognition in the HCA prototype.*
>
> *4. Tackle as soon as possible the potential robustness problem of the NLU (especially for HCA prototype) when associated with speech recognition.*
>
> *Actions required.*
>
> *Item 3 and 4 concern NISLab:*
>
> *Intensify your efforts and report on the approaches and results in the final review. Provide an intermediate progress report on these two issues at the latest by 1 October 2004.*

It must be emphasised that, despite the text of the request above and what it might be interpreted to suggest, (i) all points mentioned in the text form part of the contractual commitments of NISLab according to the NICE contract, and (ii) NISLab is not behind the plans in the contract in any way nor was NISLab behind schedule by the time of the review in May 2004. On the contrary, the natural language understanding module presented at the review was 8 months ahead of schedule. Speech recognition integration could only start around 1 July 2004, i.e. more than a month after the review, because it was only then that we, in conformance with the Work Plan, received the trained recogniser from partner Scansoft. Speech synthesis integration had been demonstrated twice by the time of the review and we explicitly stated at the review that we planned to identify and integrate the best synthesiser for our purposes in due course before completing the second Andersen prototype. The natural language understanding (NLU) robustness problem could only start to be addressed when we had available an integrated trained recogniser and had done substantial work on recogniser vocabulary and language model development. For all these tasks, we have simply followed our development plan.

The present report provides a brief overview of the present status, as of end September 2004, of planned development at NISLab as regards speech synthesis, speech recognition, and analysis of NLU robustness. Since the plan involves a number of important innovations relative to the first Andersen prototype presented at the review (PT1), these innovations and their state of development are presented very briefly in Sections 2 and 3. Otherwise, if will not be possible to understand the nature of the recogniser and NLU tests reported below. Section 4 addresses the status of state-of-the-art speech synthesis. Section 5 addresses the status of state-of-the-art speech recognition. Section 6 describes recent tests of the recogniser. Section 7 describes a test of the NLU's robustness to recogniser error. Appendix 1 shows the test corpus used in the recogniser test described in Section 6.2. Appendix 2 shows the test corpus used in the NLU robustness test described in Section 7.

2 Relevant innovations in brief

Compared to PT1, a more sophisticated natural language processing approach has been adopted for the second prototype (PT2). The approach is being implemented and tested at the time of writing. The natural language understanding module (the NLU) now forwards *domain* *ontology-based concepts* to the character module, i.e. semantic expressions which reflect an ontological analysis of the user's spoken input relative to Andersen's knowledge domains. To analyse the NLU concepts at run-time, a new module has been developed from scratch for the character module, called a Conversation Mover, or Cmover. The task of the Cmover is to analyse the received NLU concepts and either return *no move* in cases in which the input does not meet the minimum requirements for matching any output, one or several *output labels,* reflecting successful match(es) with the *key semantics* for each output, or *overdetermination* in cases where the input includes concepts which go beyond Andersen's domain knowledge. The Cmover returns are sent to the conversation intention planner which then uses its plans, knowledge of the domain ontologies, and knowledge of the discourse context to determine Andersen's next output. This new PT2 approach to input understanding is well suited for automatic generation of flexible responses for all manner of flawed or out-of-domain input, as well as for brief, to-the-point automatic response generation in general.

3 Development status

The new NLU approach was completed in mid-July 2004. By 1 August 2004, all Andersen's knowledge domains had been re-designed based on domain ontologies, and a supra-domain design specification had been completed as well.

Based on these results, rapid NLU, Cmover, and conversation intention planner prototyping has been going on, focusing on the Life domain. A comprehensive test has been made of the NLU and Cmover's handling of the Life domain, using typed rather than spoken input to the NLU. Analysis of the test data has been completed as well, showing that the NLU/Cmover pair worked quite well in this test. Briefly, 239 Life domain inputs from the PT2 development corpus were processed by the NLU and the Cmover. 227, or 95%, of these were processed successfully. Thus, we are confident that the new NLU and Cmover technologies will be adequate for PT2.

In parallel, we have developed PT2 speech recognition. However, the two development streams of NLU/Cmover, on the one hand, and speech recognition, on the other, are not working in tandem right now. Since the development data for the NLU/Cmover test just described has not yet been included in recogniser language modelling (cf. below), and since the Cmover is still in development as regards the other Andersen domains covered by our latest language model, it has not been possible before 1 October 2004 to test the speech recogniser-NLU-Cmover "chain" with data which was *both* (i) used to train the recogniser's language model and (ii) was likely to be generally familiar to the new Cmover. We *could,* of course, had searched by hand, as it were, the presently prepared language model corpora for input utterances which were also amenable to processing of the current Cmover. However, there is precious little data on the Life domain in those corpora, a main reason being the highly Andersen-driven conversation on his life in PT1.

The following data on the performance of the recogniser and of the NLU, and the approaches adopted to gathering this data, must be interpreted in the light of the current status of module development described above.

4 Speech synthesis

We have integrated state-of-the-art speech synthesis for the second Andersen prototype. Synthesis is done by the male UK English voice from AT&T which we found was the best voice for the purpose on the market.

5 Speech recognition

We have integrated state-of-the-art speech recognition for the second Andersen prototype. Recognition is done by Scansoft's speech recogniser.

5.1 Acoustic modelling status

The basic acoustic model of the recogniser has been trained by Scansoft with NISLab's Wizard of Oz 1 and Wizard of Oz 2 data as well as on a relatively small amount of internal Scansoft data. Training with more NISLab data is in progress at Scansoft.

5.2 Vocabulary and language modelling status

Iterative language modelling and recogniser vocabulary development is being done at NISLab at the time of writing. The current status is the following:

The recogniser currently works with *VOC4*, i.e. our fourth PT2 vocabulary version. VOC4 includes 1888 words (or word forms) and is based on careful analysis of data from: our WoZ1, WoZ2, WoZ3 (user test) corpora, our first prototype development corpus, and various special word lists created for PT2 purposes. PT2 will include VOC5 which will be based on additional development corpus data and is estimated to include 2000+ words.

Two language models have been developed so far. The first, *LM1*, is based on the user test corpus (WoZ3), the second, *LM2*, on LM1 + the WoZ1 corpus. Three additional language models are in development together with various experimental sub-language models. The first one, LM3, will include LM2 + the WoZ2 corpus; LM4 will include LM3 + the PT1 development corpus; and LM5, some version of which will be included in PT2, will include LM4 + the PT2 development corpus. Referring to Section 3 above, it will only be at this point that the speech recogniser, the NLU, and the Cmover "speak (exactly) the same language", ensuring that the speech recogniser vocabulary and the NLU lexicon are identical, and ensuring that the recogniser and the Cmover have been trained on the same corpora.

Given the discrepancy between recogniser training and NLU/Cmover development described above, we have performed two different types of test for the purposes of the present report. Two tests, described in Section 6, have focused on stand-alone recogniser performance, and one test, described in Section 7, has focused on NLU robustness.

6 Testing the recogniser

This section describes two tests of the performance of the recogniser in stand-alone mode.

6.1 First recogniser test

A first, simple test was performed in an office environment with a single speaker (the present author), 25 utterances from outside the training data, and a 1600 words recogniser vocabulary (VOC3). The speaker had the opportunity to inspect the recogniser output. If the output was poor, the speaker had a second chance to speak the utterance. The best recogniser output of the two was chosen for the statistics.

The recogniser was tested in two conditions, (i) without any language model and (ii) with LM1. The word correctness rate of the recogniser without language model was 36.8%. When adding our first language model, the word correctness rate rose to 88.8%.

The immediate conclusion on this first test of Scansoft's trained recogniser was that we are on the right track with our language modelling work and that the coming months should be spent on recogniser optimisation rather than on basic problem solving in order to make the recogniser work at all.

6.2 Second recogniser test

The second test was performed in a sound-proof room by two adult speakers other than the present author, one female and one male, using a test corpus of 60 utterances from the LM2 training data, an 1888 words vocabulary (VOC4), and LM2. The speakers spoke each utterance once.

The table below shows the results. 60% of the input were flawlessly recognised and 40% had various kinds of error (false deletions, substitutions, false insertions). The recogniser errors were then inspected by two researchers involved in NLU development and testing in order to estimate how many of the recogniser errors were non-fatal and fatal, respectively, as regards subsequent NLU recovery from those errors. The estimate, shown in the table, is that the combined recogniser/NLU understanding success of the 120 test utterances lies at 87%.

	SR correct	SR errors all	SR errors non-fatal	SR errors fatal	SR/NLU expected success
User1	36 = 60%	24 = 40%	18 = 30%	6 = 10%	54 = 90%
User2	36 = 60%	24 = 40%	14 = 23%	10 = 17%	50 = 83%
Total	**72 = 60%**	**48 = 40%**	**32 = 27%**	**16 = 13%**	**104 = 87%**

Since, in real-life use of the system, we cannot assume a sound-proof room, good-quality speakers, such as those doing the test, nor fully within-LM input, this result shows that further LM optimisation is mandatory. Still, the results may be described as encouraging and very much in line with the result of the first recogniser test described above.

7 Natural language understanding robustness

7.1 An approach to NLU robustness testing

In order to address the issue of NLU robustness before 1 October 2004 and for the sole purpose of the present report, the following approach has been adopted.

We have chosen 50 input utterances from the Life domain development corpus for PT2. For these utterances, we know that the NLU can analyse them correctly and we know that the Cmover will be able to identify correct output labels in all cases. In other words, for each input *as uttered* by a test subject, we can provide the correct NLU semantics and predict the Cmover's output based on this semantics. This implies that, when the speech recogniser is added to the setup, and if the Cmover eventually does not identify the predicted output, then we know that the error *must have been caused* by the recogniser.

Moreover, given the fact that recognisers produce both smaller and more insignificant errors and larger, massive errors, it becomes possible to judge the NLU's robustness in the following way: we look at all the cases in which the recogniser produces errors, however large or small. We then look at the percentage of those cases in which the NLU, nevertheless, manages to provide the Cmover with information sufficient for identifying the correct output label. This percentage will provide a first indication of the NLU's robustness. Secondly, two researchers involved in NLU development and testing estimate the current under-performance of the NLU, if any, in order to assess how much more of the input will be robustly processable by the NLU for PT2. This provides a measure of the NLU's current *lack* of robustness as well as setting a specific target for further NLU development.

Of course, as a result of the development discrepancy described in Section 3, the test just described is somewhat artificial in the sense that the recogniser does not yet have the language model it needs in order to stand a good chance of recognising the input. For this reason, we must expect relatively low recogniser correctness in the test, cf. the difference between 36.8% (without any language model) and 88.8% (with LM1) reported in Section 6.1 above.

7.2 First NLU robustness test

4 speakers, two female and two male, three from Denmark and one from Italy, each spoke the same 50 input sentences to the SR-NLU-Cmover system in a sound-proof room. Each sentence was spoken only once. All sentences were perfectly understandable to the NLU and the Cmover was perfectly able to process their semantic representation.

Due to the lack of adequate language modelling, the recogniser made a large number of errors.

7.2.1 Errors propagated to the Cmover

Looking first at how the recogniser errors were propagated through the Cmover, resulting in wrong output label selection, we found that the Cmover produced correct output for 56,5% of the 200 user inputs, ranging from 64% for the best speaker to 48% for the least successful speaker. In % of the 50 sentences spoken by all users, we found:

- no correct output for any speaker: 6 inputs = 12%.
- correct output for a single speaker: 7 inputs = 14%
- correct output for two speakers: 14 inputs = 28%
- correct output for three speakers: 14 inputs = 28%
- correct output for four speakers: 9 inputs = 18%

The data underlying these figures has not been fully analysed but the 12% no correct Cmover output for any speaker was due to the fact that important words were either lacking in the recogniser's vocabulary *(shoes, memories, apartment, early, finance, career)* or doubly present in the recogniser's vocabulary but not so in the NLU's vocabulary. The example is the word *mum.* The recogniser had both *mum* and *mom* but the NLU had only *mom.* More detailed analysis will no doubt find that, in many cases of multi-user error on a particular utterance, the recogniser's insufficient language model is the culprit.

7.2.2 NLU recovery

In the table below, U(n) is the subject id. The table shows the percentages of cases in which the NLU recovered from recogniser errors, eventually enabling the Cmover to identify the correct output label. Simply put, the Cmover identified correct output for 113 in 200 subject utterances(56.5%). Of these 113 cases, only 60 were ones in which the recogniser recognised exactly what the subjects said. In the 53 additional cases of Cmover success, the NLU recovered from recogniser error in such a way that the Cmover identified the correct output label.

It is worth noting the header of Column 5 from the left, i.e., "Minimum NLU/Cmover success". It is a success for the NLU/Cmover to correctly identify the output corresponding to the subject's spoken input. However, the NLU/Cmover success may be somewhat higher than the 56.5% stated in Column 5, simply because, given the *recogniser's output* rather than what the user actually said, the NLU/Cmover may well have made additional correct choices. However, this analysis is not so important to the present report.

	SR correct	**SR errors all**	**NLU/ Cmover recovers**	**Min. NLU/ Cmover success**	**NLU/Cmover recovery possible**	**Total recovery**	**SR fatal error**
U1	22 = 44%	28 =56%	10 = 20%	32 = 64%	6 = 12%	38 = 76%	12 = 24%
U2	10 = 20%	40 = 80 %	18 = 36%	28 = 56%	8 = 16%	36 = 72%	14 = 28%
U3	16 = 32 %	34 = 68 %	13 = 26%	29 = 58%	9 = 18%	38 = 76%	12 = 24%
U4	12 = 24 %	38 = 76 %	12 = 24%	24 = 48%	10 = 20%	34 = 68%	16 = 32%
Sum	60 = 30%	140 = 70%	53 = 26.5%	113 = 56.5%	33 = 16.5%	146 = 73%	54 = 27%

Having looked at NLU robustness above, we also need to ask about the extent to which the NLU could be further optimised. This is done in Column 6 from the left in the table above. It appears that we have room to further optimise NLU robustness by 16.5%, yielding a total NLU understanding success of the user's input in the test of 73%. This figure is 10% lower than the 87% found in Section 6.2, no doubt due to (i) LM2's lack of correspondence with the NLU and the Cmover and (ii) the recogniser's lack of certain important input words in its vocabulary. In fact, the number of system errors involving inputs with missing vocabulary words was 20, or 10%. In 12 input cases (6%), the NLU successfully recovered from recogniser errors due to inadequate recogniser vocabulary.

7.2.3 Examples

First an example of fully correct recognition, NLU processing, and Cmover output:

User: please tell me about your youth in copenhagen

Recogniser: please tell me about your youth in copenhagen

Sentence Confidence Score : 668

Concept Recognizer Output: <dialogue_act:request> <dialogue_act_type:listen> <property:no_value> <concept:lifetime> <sub_concept:youth> <concept:location> <sub_concept:copenhagen>

Next Conversational Move :copenhagen_story

This is an example of NLU recovery from recogniser error:

User: what happened in copenhagen

Recogniser: what have and in copenhagen

Sentence Confidence Score : 309

Concept Recognizer Output: <property:no_value> <concept:location> <sub_concept:copenhagen> <dialogue_act:question> <dialogue_act_type:general>

Next Conversational Move :copenhagen_story

This is an example of fatally failed NLU recovery from recogniser error (career is not in the recogniser's vocabulary):

User: i want to know about your author career

Recogniser: i want to know about your often korea

Sentence Confidence Score : 447

Concept Recognizer Output: <country:korea> <dialogue_act:question> <dialogue_act_type:general>

No Move Found

Finally, this is an example of a failed NLU recovery from recogniser error, which can be solved through NLU optimisation:

User: tell about your life

Recogniser: tell about your life who

Sentence Confidence Score : 597

Concept Recognizer Output: <user_intent:listen> <dialogue_act:question> <dialogue_act_type:person> <property:no_value> <concept:life> <sub_concept:general>

No Move Found

8 Appendix 1: Recogniser test corpus

how old are you
i come from denmark
my name is
what is your favourite fairytale
excuse me
can you tell me about one of your fairytales
tell me about the little mermaid
what do you like to play
can you tell me the morale for the fairytale the princess and the pea
tell me about your dad
what do you know about
what can you do
what other stories did you write
where do you live
what is your favourite game
no
yes
i am fifteen years old
i am a girl
what is your name
who are you
can you tell me about your grandmother
hello are you there
yes please
did you have many friends in school
would you tell me a fairytale
what is your best fairytale
which games do you like
what do you like to play
what are you doing now
yes i do
no can you tell me
it is all right i think
i like it
it is great
can you tell more
i think it is true
what is the greatest history you have made
i did not say anything
i think the ugly duckling
i read a lot of them

would you mind tell me another fairytale
could you tell me about your father and your mother
i think it is a fantastic fairytale
what did your father do
how many fairytales have you written
how are you
nice to meet you too
how tall are you
it is very nice i think it is very beautiful
that is okay did you have any brothers or sisters
i do like your fairytales
i do not know that one will you tell me about it
well he is nice
harry potter
see you some other time
i would like to know what your name is
how old were you when you died
you like to write poems
could we talk about something else

9 Appendix 1: NLU robustness test corpus

tell about your life
how was your life
your life what can you say about it
when did you live
did you live from eighteen hundred ten to eighteen hundred eighty
i think you lived in the old days
when were you born
you must be born more than a hundred years ago right
your birthday
you died when
did you die in eighteen hundred ninety
you died in eighteen hundred seventyfive
was your family rich
what about your family
do you also have a brother
what did your parents do for a living
what did your dad do
didnt your dad repair shoes
what do you remember about your father
what did your mother do for a living
what do you remember about your mom
how about your granddad
didnt your grandfather die when you were very young
tell me about your grandmother
didnt your grandmother die when you were a little
how was it when you were a little
which memories do you have of your childhood
were you a nice boy
where did your family live
whats the name of the city where you were born
i know that you come from odense
how was your childhood home
how big was your childhood apartment
did you have your own room as a child
were you good at school
did you go to school when you were a boy
which memories do you have from your school years
which kind of games did you play
what is your favourite game
i like to play do you
how was your youth

i want to know something about your teenage time
why did you go to copenhagen
why did you go to copenhagen so early
what happened in copenhagen
how did you finance your first years in copenhagen
how was it to live in copenhagen
please tell me about your youth in copenhagen
please tell me about your life as a grown up
i want to know about your author career