



VoiceWeb: Wizard of Oz

Annelies de Bruine
Internet and Mobile Systems Laboratory
HP Laboratories Bristol
HPL-2001-47
March 16th, 2001*

multimodal,
wizard of oz,
speech, user
studies,
usability

The VoiceWeb project works on spoken language technology. The speech system we work with is the Galaxy system of MIT, it is a speaker independent system. Like most speaker independent speech systems, Galaxy works with different domains. Every domain has its own recogniser and will be able to understand words and sentence within a certain field, the domain. The VoiceWeb team in HPLabs Bristol works on the so-called MIZIK-system. The domain is music and at the moment of the trials the database held information about the album top 40. To train the system we needed to collect utterances. The Wizard of Oz experiment was designed to collect words and utterances. People could ring a free 0800-number and ask our system 10 questions. The callers thought the system was answering the questions, but there was an operator listening in who gave the answers. The operator listened to the questions, typed in a code and the answer was given to the callers by the system via a speech-synthesizer. Over 300 people called the 0800-number and their utterances have been used to upgrade the MIZIK-system.

Introduction & Method

The VoiceWeb team in the research labs of Hewlett-Packard in Bristol is working spoken language technologies. We are working with the GALAXY-system, developed by MIT. At HPLabs we are working on the GALAXY-system in a music domain. To train the system we need to collect English utterances within our domain.

From the 20th of July until the 4th of August 2000 we run an experiment to collect utterances to train our system. To collect these we set up a call centre in the lab and opened up a free 0800-number. Via different media we asked people to call our number and ask our system 10 questions about music.

The call centre was manned for 12 days. Weekdays only from 9 am until 6 pm. One person who also acted as ‘wizard’ manned the call centre and there were always people of the VoiceWeb team to take over if necessary and to help out in case the system would crash.

To encourage people to call our number, we decided it had to be a free telephone number. The free phone number was advertised on a regional radio station (Galaxy101), in the music section of Bristol’s local newspaper (Full On in the Evening Post), flyers were handed out during events, in pubs and in clubs and we send e-mails to friends and family. In the adverts we asked people to call the free phone number, ask an automated system 10 questions and if they managed to ask our system 10 questions they would receive a free cinema ticket. We purchased 500 Warner Village cinema tickets to give away to the callers. In the adverts it was stated that only the first 500 callers would receive a ticket.

The calls were recorded and later transcribed. All utterances were recorded as .wav files. The VoiceWeb team will use the .wav files and the transcriptions to train the system.

Close to 300 people called in the 12 days the lines were open. When people called they first got a message that the call would be recorded for research into spoken language systems and that all music information was fictitious.

After that the caller got through to an operator who explained the procedure and took some personal details for research purposes and name and address details were taken for sending the cinema tickets. The name-and-address files won’t be kept on record. After the explanation the caller got transferred to our automated system. For the script for the operator, see Appendix A.

The system first welcomed the caller and then asked the caller to ask a question about music after the beep. The caller asked a question and the system tried to answer it. The operator also functioned as a so-called Wizard, named after the Wizard of Oz. The operator listened in and tried to answer the questions by typing in short codes. Callers thought the system answered the questions, but it was the wizard typing in codes to give answers. Most questions couldn’t be answered by our system because we didn’t have the information asked for in the database. The system only held (fictitious) information about the, at that time, current album top 40. If the system couldn’t answer the question, the wizard typed in a ‘sorry-message’ by typing in `ss#`. `Ss2` stood for ‘say sorry message 2’, which was: “Sorry, I am not able to answer your question. Please ask another question about music after the beep”.

Here is a list with the standard answers the wizard could give:

1. Type in the welcome message, `s ml`

Welcome to the music service! Please ask a question about music after the beep and wait for a reply

2. If you can't answer the question: [s s2](#)

Sorry I cannot answer that question. Please ask another question about music after the beep

3. If you can't answer the question: [s s3](#)

I do not have any information on this question. Is there anything else I can do for you?

4. Answer the question or [s m4](#)

Thank you, but I can only answer specific questions relating to the current album top 40. You can ask about artists, prices, chart positions and reviews. Please try again.

5. Answer the question or [s s5](#)

Sorry, can you please ask another question about the current album top 40?

6. Answer the question or [s s6](#)

I know about artists, prices, chart positions and reviews. Go ahead and ask another question.

7. Answer the question or [s s7](#)

Sorry, I am not able to answer that question. Please try another question.

8. Answer the question or [s m8](#)

I apologise for not being able to answer your questions, but your questions help me to get better. Can you please ask me three more questions?

9. Answer the question or [s s8](#)

I only know about artists, prices, chart positions and reviews. Please ask another question.

10. Answer the question or [s s9](#)

I am sorry, but I do not have information about this. Can you ask your final question please?

11. Goodbye message [s m10](#)

That will be all. Thank you for calling. We will send you the cinema ticket as soon as possible. Good bye. [Hg](#)

There were also codes for when people didn't know what to say or when they spoke before the beep.

When the system was able to answer the question, there were codes to use for getting the title, review, price or getting the chart position for the album the caller asked for.

Type	Action
gt <artist>	Get the title of the album by the artist
gt <pos>	Get the title of the album at the specified chart position
gr <artist>	Get the review for the album by the artist
gr <pos>	Get the review of the album at the specified chart position
gp <artist>	Get the price of the album by the artist
gp <pos>	Get the price of the album at the specified chart position

We told callers that they could ask any question about music, so they could ask whatever they wanted, since we wanted to collect some data on what callers would be interested in. After the first 3 questions the system prompted it only could answer specific questions about the current album 40. At the end the system thanked the called and the wizard hang up the system.

Results

In the first 12 days 277 people called and asked our system (approximately) 10 questions. On the final 2 days (10/11 August) 91 people called. The call centre we set up in the psychology lab worked well and the operator/wizard we hired via an employment agency did a very good job.

We entered information of 355 people in our database. Most people called from a landline (270). Only person called from a car phone, 15 people called from a public telephone, 25 people used a cordless phone to call our telephone number and 39 callers used a cordless phone.

The average age of the callers was just a little bit over 23. 197 men called (55%) and 158 women (45%). Most callers were British, only 20 callers were not.

Most callers who got through called after hearing our advert on the radio station.

How Did People Find Out About The Project?

		Cost	Incl. VAT	PPC
Newspaper	39	£384.30	£409.26	£10.49
Radio	106	£3,440.00	£4,042.00	£38.13
Flyer	27	£160.00	£188.00	£6.96
Word Of Mouth	51		(ppc = price per caller)	
Email	41			
Other	7			
Unknown	6			

Total Number Of Calls: 277 From 20 July until the 4th of August

	10/11 August	Totals.
Newspaper	13	52 (15%)
Radio	*	106 (29%)
Flyer	4	31 (9%)
Word Of Mouth	28	79 (21%)
Email	32	73 (21%)
Other	2	9 (3%)
Unknown	*	6 (2%)

Total Number Of Calls: 79 356

The radio adverts had the best effect, a lot of people called after hearing our ad on the radio. Unfortunately we weren't able to handle all incoming calls (up to 54 calls in half an hour after a radio advert while we could only handle approx 4 calls (7 max.) per half hour because most people called immediately after hearing the ad.

Dates	Peaks times of the day!	Times of the radio add approx
17/07/00	16:00=8	-
18/07/00	18:00=2	-
19/07/00	No Records!	-
20/07/00	9:30=24	4:20, 7:32, 10:50, 16:52, 18:49, 23:20
21/07/00	11:00&14:00=21	4:20, 5:24, 11:20, 13:51, 14:29, 22:29
22/07/00	Weekend	Weekend
23/07/00	Weekend	Weekend
24/07/00	12:30&13:00=19	1:53, 5:20, 12:52, 13:52, 17:48, 23:24
25/07/00	16:30&15:00=10	0:53, 7:39, 12:52, 16:50, 19:24, 22:53
26/07/00	No Records!	4:53, 5:51, 6:08, 13:51, 20:24, 21:52
27/07/00	No Records!	7:20, 11:49, 13:25, 19:51, 21:25, 23:20
28/07/00	15:00=13	1:24, 3:53, 5:50, 8:26, 10:39, 15:17
29/07/00	19:30=3	Weekend
30/07/00	14:00=2	Weekend
31/07/00	17:30&18:00=9	2:40, 4:25, 8:33, 11:35, 14:40, 17:38
01/08/00	16:30=6	2:40, 7:53, 10:52, 10:52, 10:52, 23:40
02/08/00	09:00=16, 12:30&13:00=12+16, 15:30-18:00=49	1:40, 4:25, 5:42, 12:53, 15:53, 18:20
03/08/00	14:00&14:30=13+10, 17:30=22	2:24, 4:24, 8:53, 12:23, 15:10, 20:22
04/08/00	13:30=45, 16:30=23	4:40, 5:40, 9:40, 13:20, 19:40, 22:17

Dates	Calls Accepted	Calls Answered
17/07/00	27	10
18/07/00	3	0
19/07/00	No Records!	No Records!
20/07/00	88	36
21/07/00	101	41
22/07/00	Weekend	Weekend
23/07/00	Weekend	Weekend
24/07/00	87	39
25/07/00	48	27
26/07/00	No Records!	No Records!
27/07/00	No Records!	No Records!
28/07/00	63	22
29/07/00	8	Weekend
30/07/00	7	Weekend
31/07/00	49	15
01/08/00	38	15
02/08/00	150	44
03/08/00	113	39
04/08/00	195	50

TOTALS:

977

338

10/8/00	232	47
11/8/00	222	44
TOTALS:	1431	429

We received far more phone calls than we could handle. We had very quiet times and there were times that too many people called for our operator/wizard to answer. Especially after the radio advert was broadcasted we could see peaks in the call logs.

On the **10 and 11th of August** we received a lot of phone calls, it were the top days with most calls accepted and most calls answered. We put another advert in the local newspaper, which gave us 13 calls out of the total of 79. Most calls however came from people who received an e-mail (32 callers) or people who heard about it from other people (28 callers). We managed to answer 91 phone calls in these two days. Again there were peak times. On the Thursday 66 people called our phone number at 14:30 and up to 91 people tried to call in around 5 o' clock on the Friday.

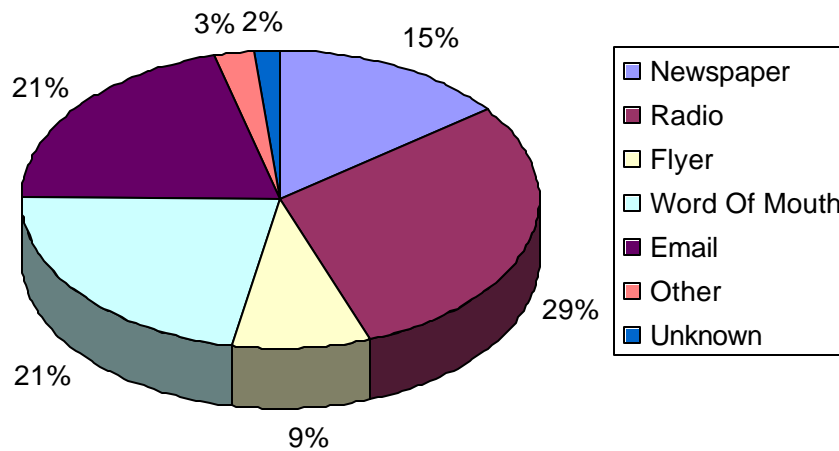
The newspaper advert got in quite a few callers. We advertised in the Thursday paper since it has a) a music section called Full On and b) it is the most sold paper in the week because it has a section with jobs vacancies. The first 2 adverts went in on Thursday 20 and 27 July. We received phone calls of people who saw the ad in the newspapers, but we also got remarks that people didn't call because they thought the 500 tickets would be gone by the time they thought about calling. For the final week we decided to change the ad to make it more attractive for people to call. We got a lot of phone calls in on the Thursday & Friday with the new advert so we decided to place another advert in the paper on Thursday August 10th. We opened the HP Music Service for two more days Thursday 10 & Friday 11 August to collect more utterances (no results of these 2 days yet).

Advert 1: 20 + 27 July	Advert 2: 3 + 10 August
<p>Call the Music Line on 0800 – 0963369 and get a free cinema ticket for Warner Village Lines are open from 20 July until the 4th of August from 9 am until 6 pm.</p> <p>Can you think of 10 questions to ask an automated music shop? If so, and you'd like to receive a cinema ticket, call the number above. An operator will give you instructions. The call is free of charge and will take a maximum of 10 minutes. The call will be recorded. We will ask your details to send you a Warner Village cinema ticket. These details won't be kept on file. The first 500 callers will receive a ticket, only one ticket per person.</p>	<p>Call the Music Line on 0800 – 0963369 free cinema ticket for Warner Village for everybody who calls today or tomorrow between 9 am and 6 pm and finishes the task.</p> <p>Can you think of 10 questions to ask an automated music shop? Call Thursday or Friday and if you ask our system 10 questions about music you receive a cinema ticket. An operator will give you instructions. The call is free of charge and will take 6 minutes. The call will be recorded, your personal details won't be kept on file. Only one ticket per person.</p>

41 people called in after getting an e-mail. We send a few e-mails when it was quiet at the call centre to distribution in the Bristol area. Quite a few people rang after receiving the e-mail message. For Thursday 10 & Friday 11 August we sent out mail messages to HP-distribution list asking people to call. We sent the message to 5 different distribution lists at different times trying to avoid clogging up the phone system; 9am – 10am – 11am – 12noon – 1pm. The e-mails worked very well, but there are only so many e-mails we could send without spamming people.

The flyers worked, but didn't attract as many callers as we hoped for. We noticed while handing out the flyers that people found it hard to believe that there wasn't a 'catch'. We handed out flyers at events in Bristol and in pubs & clubs, mainly at times the telephone line was closed. Maybe we should have started distributing during hours that the phone line was manned. Also we could have started a bit earlier with distributing the flyers, since a lot of people called after the official closing date (34 on Tuesday the 8th of August). The flyers were relatively cheap though. 10% of callers got the phone number of a flyer, which is a good score.

Which Style Of Advertising Was The Most Successful? (incl. 10/11 August)



All incoming phone calls were logged. In the logs we could see how many people called and how many phone calls we managed to answer. In the first days we saw that people hung up while hearing the opening message (This call will be recorded. All information about music is fictitious). We don't know whether people thought the line was busy when hearing the message or whether they weren't interested in pursuing with the call being recorded or that they didn't want to listen to fictitious information.

Some people hung up after hearing the message that the system was only able to answer questions about the current album top 40. Other people got annoyed because the system wasn't able to answer their questions. We added an extra sentence in the explanation about the importance of collecting utterances for our research and asked people to please not hang up but continue asking questions, even when the system wasn't able to give a proper answer. We also changed a few messages (like adding sm8 to the list) to explain that we needed the questions to improve the system ('your questions help me to get better').

We wrote down all incoming calls and times on a piece of paper next to the telephone with remarks if something remarkable happened either because of the system or the caller. Like we wrote down whether people hung up halfway or when the system showed a little fault.

The system sometimes had trouble with callers pausing or hesitating (uhhm). After a pause or after saying 'uhhm' the system gave the beep for end of utterance, which also meant end of recording. The operator/wizard discussed this with the VoiceWeb team and it did improve after changing something in the system. We added an sm7, 'go ahead' to encourage people to go on and solve the problem with the beep after 'uhhm'.

The database didn't always work: We couldn't get the title for the Abba album while it was in our chart (gt abba) or the title for the Westlife album (gt westlife). Also it didn't work with underscores when callers asked about Whitney Houston or Britney Spears (whitney_houston, britney_spears). Chart positions were hard or impossible to get as well. And asking for S Club 7 gave "Apostrophe S Club 7 feet". Also the reviews could have been written a bit better. We didn't have time before the experiment started to write proper reviews since it is very time consuming work.

Most callers did end their session and managed to ask the system 10 questions. Overall it was an interesting experience. The call centre worked very well and the system only let us down on the first day. The system got very slow and didn't record all utterances anymore. The problem was solved very quickly.

After talking to some friends and family about the experiment, we noticed that quite a few people said they didn't call because they thought the 500 tickets would have been gone within the first few days. For legal reasons we had to let people know in our advertisements that there were only 500 tickets to give away, even though we thought we wouldn't be able to handle more than 500 calls in 12 days. We lost a few callers because of that, but we thought it would be better to be safe than sorry.

Summary and Discussion/Conclusion

The experiment with the call centre was a success, even though we were hoping for a few more phone calls, which is why we opened up the lines for an extra 2 days. The VoiceWeb team now has listen to the recorded utterances to find out what the quality is of the recordings.

Advertising our free telephone number went well. The cinema tickets were an extra encouragement for people to call. The radio advert got a lot of reactions, too much to handle sometimes. People called our phone line directly after they heard the message. If we repeat this experiment it might be a good idea to work with an extra operator/wizard after the advert has been on air. 14% of the callers found the telephone number in the newspaper. Especially after changing the ad (from max. 500 tickets to everybody who completes the task) we got more response. The flyers got 10% of the total amount of callers in. E-mails worked very well, but it is hard to decide when sending e-mails to distribution lists is spam. Spreading the e-mails over the day worked well to stop clogging up the telephone line.

On the 10th and 11th we had many incoming phone calls of people who either got an e-mail or via word of mouth. We sent the e-mails to distribution lists within Hewlett-Packard Labs in Bristol. The big response might have something to do with the fact that people in the labs are used to invitations to participate in research. Even though we did spread the e-mails, a lot of people tried to call in at the same time. Especially Friday after 5 o'clock it got really busy. Probably people who finished work and tried to call before going home.

The database held fictitious information about the current album top 40. The information was not complete and the short codes didn't always work. For the experiment it worked fine though. Only a few people got that annoyed that they hung up and the frustration of the system not being able to answer most questions was resolved by some extra explanation in the introduction by the operator and to add another 'sorry message'.

The utterances of both the wizard and the callers were recorded (.wav files). Sometimes we didn't hear the second beep, which meant end of recording. One person called from a car phone and only one sentence was recorded. This might be because of too much background noise. It is possible that the system didn't recognise the beginning and/or end of a sentence because there was a lot of noise and therefore only recorded the first utterance.

The members of the VoiceWeb team should now evaluate the quality of the data. After we have these results we can start planning our future research into spoken language systems.

Further research

The next study will be study into usability and user satisfaction. People who called our free phone number and completed the task received the cinema ticket and a letter. In the letter people are invited to take part in further research. A similar experiment will be set up in the labs without using a call centre. Next time we will focus on the experiences of the users as well as collecting more utterances and we will invite subjects to come to our lab.

To collect more utterances we can also keep the line open but without an operator/wizard, completely automated (and without sending out cinema tickets). We need to think about how to

set this up and what we want the callers to do. We can start with a simple task and gradually make the tasks for the callers more complicated when the system improves.

To improve the dialogue it would be good to record proper dialogues so we can teach the system about contexts. We might try to set up a small trial to collect this information. We might use directed dialogue for this trial.

It will be a good idea to collect real data. For that we need a better database. We can collaborate with other partners who have databases set up and updated daily. We can work with a radio station or maybe with an (online) record store. It would be good to collect real data and to work with real information.

The utterances we collected can be used to find out what the people are interested in; what kind of information do people want when calling a music service. We can take a look at the first 3 questions the callers asked, before they got prompted that the system only held information about the current album top 40. We can improve our system by analysing the questions asked by the callers.

After that we can move to domain research. Would people stick to the domain (music) or do people want more out of the system? We can have a look at the results of this experiment to find out what people see at a music domain by looking at the first 3 questions the callers asked. We can research whether people talk in domains in real life and what size of domain would be useful for a spoken language system.

Finally these results can be used as a start-up for research into multimodal systems. We can look at different scenarios and what combinations will make sense for users of multimodal systems. We can also test whether people like different ways for information input and output and what people prefer (clicking, voice control, touch screens, etc.)

Annelies de Bruine
HP Labs Bristol, August 2000

Appendices

I would like to thank all members of the VoiceWeb team at HPLabs Bristol (Paul Brittan, Marianne Hickey, Steve Hinde, Guillaume Belrose, Lawrence Wilcock, Michael McTernan, Jane Mather, Jeff Pillou, Yu Chien Chan) Erik Geelhoed for the use of his lab, MIT for the Galaxy system and Daniel John, our fantastic operator/wizard.

Appendix A

The system

Open main terminal (x terminal) by clicking the [shell](#) on the bottom tool bar

Click [[one](#)] on the bottom toolbar

You will be connected to Brittan-p-6

1. Type in `cd /home/Galaxy2.2.1/scripts` [\[return\]](#) (*space after cd*)
 2. `tcsh` [\[return\]](#)
 3. `source ../templates/GALAXY.setenv` [\[return\]](#) (*space after source, before ..*)
 4. `start_hp_live_servers.cmd`
- ? four windows will pop up: check if they all say 'listener'
5. Click yourself back into the root@Brittan-p-6 window
 6. `start_hp_live_hub.cmd`
- ? hub window will pop up: check if it works, wait for the dots to show

Click [[two](#)] on the bottom toolbar

1. Open Netscape
 2. Open the Wizard Home Page under bookmarks
- >> Your database for putting in personal details is open

Click [[three](#)] on the bottom toolbar

1. Open another Netscape
 2. Open webGalaxy under bookmarks
- >> Your wizard window is open

What if the system crashes (it won't hang up or doesn't answer anymore)

Close windows

1. GUI first
2. Then hub
3. Close others but leave Brittan-p-6 open and type in `start_hp_live_servers.cmd`
 - ? four windows will pop up`start_hp_live_hub.cmd`
 - ? hub window will pop up

For questions about the system, call Paul, Yu Chien, Guillaume or Jane

For all other questions, call Annelies, Marianne, Steve or Jean Francois

BACK UP:

Back Up from Friday 21 July

(Check if operators are OK, take over for lunch & coffee/tea breaks)

- > Paul 9-11 (* coffee break at 10.15!)
- > Annelies & Yu Chien 11-14 (Annelies only on 24, 25, 26 July)
(* Lunch from 12 -!)
- > Jane & Jean-Francois 14-16 (Yu Chien only on 24, 25, 26 July)
Michael on Friday the 21st
* Tea break at 3.15!)
- > Steve & Marianne 16-18 (Steve only on 24, 25, 26 July)

Start: 8:45

Coffee: 10:15 – 10:45

Lunch: 12:00 – 13:00

Tea: 15:15 – 15:45

Finish: 18:00

Make sure that between 9 am and 6 pm there is always somewhere in the lab to answer the phone.

The telephone:

- Log-on

To logon, press [in-calls](#)

Enter your agent ID 0501##

- Ready to take phone calls

The phone is not ready, people will get the message that the line is busy

To get ready to answer phone calls, press [in-calls](#) again

You will get the caller on the line immediately. The caller has heard the message *'Thank you for calling Music Master. This call will be recorded for research into spoken language technologies. All the information about music is fictitious'* before you picked up the phone call.

When the phone rings, press [in-calls](#) and speak immediately

Thank you for calling the HP Music Master.
I will explain the procedure to you. Is that OK?

First I will ask you a few questions about yourself. After that, I will transfer you to an automated system. You will be prompted to ask 10 questions about music and the system will try to answer your questions.

With your questions about music we can train our voice recogniser.

.... Any question about music will do

Is that clear? You think you can manage?

➤ **Only when needed you can give examples**

You can ask about reviews or ask the price for an album you'd like to buy

If clear - Get details of caller and put details in database [two]

Put details in database

I will now enter your details into our database

- (You can guess gender I presume)
- May I ask you how old you are?
- What kind of telephone are you using at the moment? Are you calling us with a landline, a cordless phone, mobile?
- Are you British?

If you manage to ask our system 10 questions, we will send you a cinema ticket. To do this, we need your name and address. We will not keep these on record

- Can I have your name please?
- And your address?
- (In which city is that?)
- Postcode?

Thank you. I will now transfer you to our automated system. The system will first welcome you and then prompt you to ask a question about music after the beep.

Are you ready? I will transfer you now. It might take a while, so please be patient.

To transfer, press [conference](#) and the [button with 22363](#)

[Mute](#) your own telephone, so people can't hear you type.

Press [conference again](#) after the system picked up the phone.

End of this database, please go to your wizard window [[three](#)] on bottom toolbar & get ready for the welcome message (don't forget to mute your phone!!)

- Change window to wizard window [[three](#)] - Netscape
 1. Type in the welcome message, [s m1](#)
 2. If you can't answer the question: [s s2](#)
 3. If you can't answer the question: [s s3](#)
 4. Then [s m4](#)
 5. Answer the question or [s s5](#)
 6. Answer the question or [s s6](#)
 7. Answer the question or [s s7](#)
 8. Answer the question or [s s8](#)
 9. Answer the question or [s s9](#)
 10. Goodbye message [s m10](#)

Answers to the questions and other messages are on the sheet in front of you.

s = sorry message

m= message

gp = get price

gt = get title

gr = get review

look for artist list on the wall in front of you

- Hanging up

If caller completed the task, type in goodbye message 3 [s m3]

Hang up the system by typing in **hg** in the wizard window

- Log-off For logging off, press **make busy 2x**