CREATING A SPONTANEOUS CONVERSATIONAL SPEECH CORPUS

Maria Husin, Darryl Stewart, Ji Ming, F J Smith*

School of Electronics, Electrical Engineering and Computer Science
Queen's University Belfast, Belfast, BT7 1NN, N. Ireland
**Email:* fj.smith@qub.ac.uk
Email: dw.stewart@qub.ac.uk, j.ming@qub.ac.uk, m.husin@qub.ac.uk

ABSTRACT

Speech recognition and language analysis of spontaneous speech arising in naturally spoken conversations are becoming the subject of much research. However, there is a shortage of spontaneous speech corpora that are freely available for academics. We therefore undertook the building of a natural conversation speech database, recording over 200 hours of conversations in English by over 600 local university students. With few exceptions, the students used their own cell phones from their own rooms or homes to speak to one another, and they were permitted to speak on any topic they chose. Although they knew that they were being recorded and that they would receive a small payment, their conversations in the corpus are probably very close to being natural and spontaneous. This paper describes a detailed case study of the problems we faced and the methods we used to make the recordings and control the collection of these social science data on a limited budget.

Keywords: Spontaneous speech, Natural conversations, Speech corpus, Human data

1 INTRODUCTION

One of the difficulties encountered frequently by research groups on statistical models of Speech (Jelinek, 1985; Young, 1996) and Natural Language (Jelinek, Mercer, & Bahl, 1983; Kuhn & De Mori, 1990) is that the statistical models developed and tested in the laboratory fail when they are tried with spontaneous speech arising in natural conversations (Furui, 2003; Furui, 2005; ACMSIR, 2007). Previous models have been designed for the recognition of read speech and can achieve a recognition rate of greater than 95% for speech read from a book or from a prepared text in laboratory conditions. But the recognition rate for conversational spontaneous speech is much lower and typically not greater than 65%, even when the training is on spontaneous speech (Furui, 2003). The problem is that the way in which people speak in conversation with one another is very different from read speech recorded in an experiment in an academic department. Spontaneous speech is often not grammatical; sentences are not finished; there are hesitations; there are interjections, such as *ah*, *uh*, *uhm*, or *er*; many words or sequences of words are repeated; the speakers interrupt one another, speak over one another, use different words (some not in normal dictionaries), different prosody, different intonations, and different accents.

There are few databases of spontaneous conversational speech available to academics apart from those recorded by the "CallHome" project of the Linguistic Data Consortium (LDC) of the University of Pennsylvania (Cieri & Lieberman, 2000). This has recorded several conversational databases in different languages, including American English. There are no large databases freely available in English and none in English from the UK or Ireland. We therefore decided to build such a corpus, a corpus of telephone conversations mainly restricted to local students, their friends, relatives, and some others at the university. It was planned that conversations would be 15 minutes long with a goal of over 200 hours of conversations to be recorded. The corpus would be named "The TitaniQ Corpus"; this name originates from the dry dock of the Titanic that lies adjacent to the offices of the authors and from the letter "Q" representing Queen's University.

In the following case study, we begin by describing our first efforts in two disappointing pilot projects to persuade students to make the recordings we needed, and then we describe the third successful procedure.

2 THE PILOT PROJECTS

2.1 In the student building

We first used our experience of a successful experiment with students a few years earlier, comparing human performance with a statistical language model (Owens, O'Boyle, McMahon, Ming, & Smith, 1997). We had then rented a room in the student union (the student building housing restaurants, shops, bars, etc.) frequented throughout the day by large numbers of students. We had used leaflets offering a payment of £5 (about $7.50) to entice students into the room during the day to take part in the experiment. This had been very successful, and we planned therefore to repeat this procedure for our database of conversations.

This time we rented a room with a normal land-line telephone and connected the phone to a recording device. We offered £5 and a free phone call for a 15 minute conversation. A student enticed by a leaflet, after signing a form agreeing to be recorded, would use our land-line to call a second person, often calling home, and the first step in the conversation would be to record the agreement of the second person to the recording of the conversation. This other person almost always agreed. We provided some current newspapers to give them possible topics for their conversations.

However, the room was not always available, and there was also a problem with noise in the very busy student union during the day, not helped by some building work then taking place. Therefore, we changed from day-time recording to the evenings. It was quieter but only because there were fewer students around, and we quickly discovered that students socializing in the evening did not want to come into our room to earn £5. Thus, although some recordings were being made, they were few in number, often only one recording in an hour. Also, when we analyzed the first recordings, we found that there was a slight continuous hum in the background. This might have been due to some electrical equipment in the building. Whatever it was, it was unacceptable; however, before identifying and eliminating the hum, we realized that we were recording too few conversations and a new location was needed.

2.2 In the Computer Science Department

In the second pilot project, we changed to making recordings during the day in a laboratory within the Computer Science Department where our own students were available and where we could make recordings of high quality and certainly without a hum. To advertise our project, we still handed out leaflets in the student union. Unfortunately, after a boost from students and staff within Computer Science, the number of students prepared to walk the 400 meters from the student union to the Computer Science building was disappointingly small; in the evenings in the student union sometimes an hour or more would pass, and no students would come.

There were other problems. When a student came into our recording laboratory in the Computer Science Department, the student could not be left alone because some accident might happen or something might go wrong with the equipment or, since we usually did not know all of the students, something might be stolen. Therefore, a member of staff or a mature graduate student had to be present during the recording. Unfortunately, a conversation by a student conducted in a university laboratory in the presence of someone else, sometimes a member of staff or a professor, is not likely to be completely natural! The main purpose of the project, to make recordings of spontaneous natural conversations, was being undermined.

Therefore, after about 2 months we realized that we had to take a different approach. One of the team made a visit to the Linguistic Data Consortium (Cieri & Lieberman, 2000) in the University of Pennsylvania in Philadelphia to look more closely at how they had recorded their "call home" conversations. This visit and the advice freely and kindly given to us led us to the procedure described in the next section, similar to that used in Philadelphia but with more restriction and more control.

3 THE SUCCESSFUL RECORDINGS

3.1 Cell phones and quality

In the early recordings described above we had discouraged our students from calling people who were using cell phones (mobile phones) because the quality of the sound would be better to a land-line than it would be to a cell-phone of unknown quality. However, we found that students often did not want to call anyone on a land-line; therefore, we had to allow them to make calls on their cell phones.

At the same time, we realized that to record natural conversations we had to allow the students to use whatever phone they wanted to use and to use it from wherever they wished to make their call, without someone else being present. Thus, if we wanted recordings of conversations that were as natural as possible, we had to compromise on high quality and allow cell phones with variable quality and variable noise backgrounds, as the students themselves chose. The result of this was that almost all of the calls in the corpus after the first few were made between two cell phones. The sound quality was not so good, but it was more realistic, and if humans can understand speech over a cell phone, so should our speech recognition systems. Initial study of the content of the calls appears to show that the content was indeed very natural and used a rich spontaneous language to cover all kinds of subjects. Therefore, this is how we decided to proceed.

3.2 Three-way conference call

To record the speech from the two participants, *we* had to rout the call through a recording device located in a quiet recording room set up in an attic in the Computer Science Department. This room had four lines installed, all accessed by the one free 0800 number so that there was no cost to the student. Each of these lines was connected to a phone that in turn was connected to an inexpensive PC with a good quality sound card. It was found that the sound card supplied with most PCs was not good enough; the quality was assessed by simply listening to test recordings and finding if the recordings were being noticeably distorted. With the good quality sound cards we used, no distortion was detected.

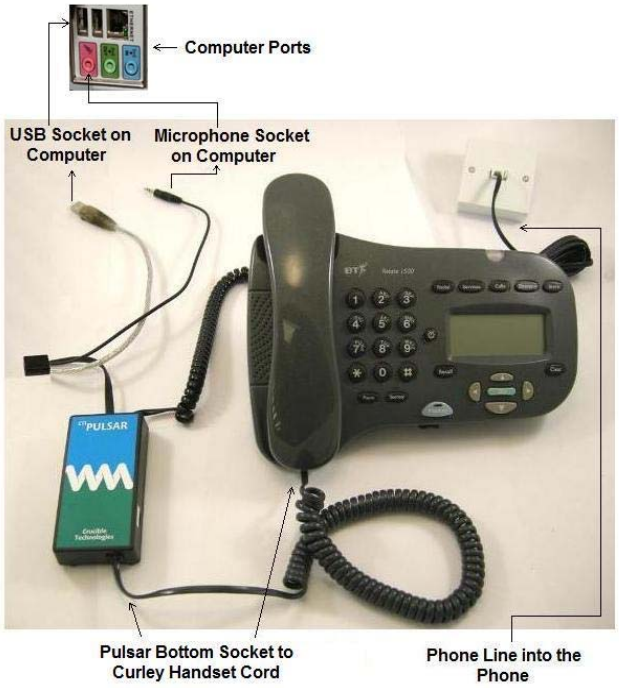

Figure 1. Pulsar Plus recording equipment, showing the break in the telephone circuit and the computer connection

The Equipment is illustrated in Figure 1. It is straightforward and needs little comment. A digital recorder (Pulsar Plus made by Crucible Technologies) was broken into the telephone circuit, recording both incoming and outgoing signals on a single channel, and sent the digital PCM data to the attached computer through a USB

connection. It was simple but effective and gave no trouble. The quality of the recording was good enough that we could not detect any distortion.

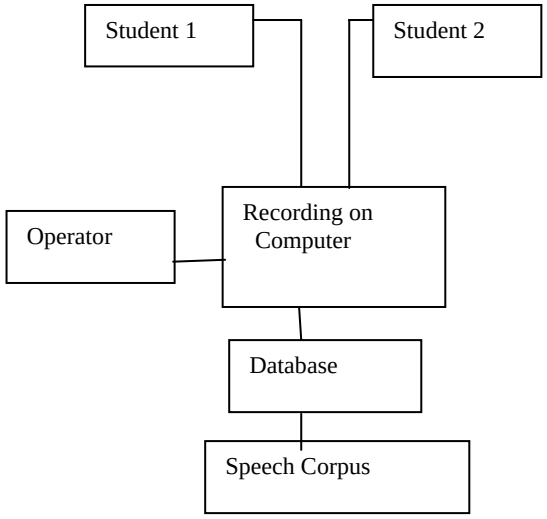

Figure 2. The architecture of the 3-way conference linking the 2 participants with the operator through the computer, which stores the data in a speech corpus, accessed via the database

The architecture of the 3-way conference system is illustrated in Figure 2. A student called the 0800 (free) number and was answered by an operator (often a graduate student employed by the project). The calling student gave the operator the number of the person to be called, and the operator then set up the 3-way conference call. The recording equipment was activated from the computer, and a digital recording of the conversation was transferred into a file, which was automatically created uniquely for that conversation. The operator checked first that the two people being recorded knew and agreed to the recording being made "for educational and research purposes". Normally conversations were for 15 minutes; after that time the operator interrupted the call, and the conversation was brought to an end. However, if the student wished to continue, and if there was no queue of students waiting to start, the operator allowed the continuation of the call for an additional 15 minutes but usually not longer.

3.3 Operator control

The process of the 3-way conference call could have been made more automatic, such as the process at the Linguistic Data Consortium in Philadelphia (Cieri & Lieberman, 2000), but the use of an operator who was employed by the project, often a graduate student, allowed a measure of control. Having an operator had several advantages. The operator could explain the process and encourage a new caller, answering questions if necessary. The operator could also listen in from time to time to ensure that a real conversation was taking place and could sort out problems on the few occasions when things went wrong. The operator also recorded metadata about the conversation, including the sexes of the people being recorded and the times and lengths of the recordings. The names and telephone numbers were also recorded on paper but not on the computer. These paper records were used to ensure that we paid the right students the correct amount of money and were then discarded. Therefore, the names and telephone numbers of the participants were not recorded in the subsequent database.

4 STUDENT PARTICIPATION

4.1 Persuading students

As is clear from Section 2, our biggest problem always was to persuade enough students and others to participate and allow us to record their conversations. Apart from initially some staff and some students in Computer Science, we relied on the incentive of the small payment of £5 for each 15 minutes of conversation. Our budget set by a small grant from the university did not allow us to spend more to obtain the target of 200 hours of conversation. The £5 was paid only to the student initiating the conversation and not to the student or other person being called, so the payment was only £2.50 for each person being recorded. We found in our first

attempts to make recordings that this would not be enough if it was not convenient for the student. The change in the methodology from recordings in the Computer Science Department to the use of cell phones from rooms or home not only made the calls more natural but also increased the number of participating students considerably. The numbers also increased because we learned how to market the project better.

Our main method of attracting students was to hand out leaflets in the student union. The early leaflets we handed out were long and full of detail on the research and the benefits to the department and university. They were probably thrown away by the students into the first wastepaper bin they saw. We quickly learned to produce more attractive leaflets emphasizing what the students wanted, i.e., the payments. We also began to use some colour and to put the detail and research objectives in small print at the bottom of the leaflets, see Figure 3. It was also found that the best time to hand out leaflets was during the lunch break in the student union. At this time there were large numbers of students visiting the union for their lunch or to use one of the union's shops. The students were also relaxed, in less of a rush, and more likely to read our leaflet. To increase our profile, we manned a desk, usually with two persons, in the entrance hall surrounded by posters between 12:30 and 2:00 PM. Some students came back to the desk after their lunch, and the project was explained to them. If they were interested, they were told a possible time when they could call the free 0800 number to our operator for the recording of their conversation.

At first we allowed recordings to be made at any time of the day. However, we soon found that few recordings were made in the mornings or afternoons; therefore, we changed to recording in the evenings only, between 6:00 and 10:00 PM. In the early evening, all 4 lines that we had available were often in use.

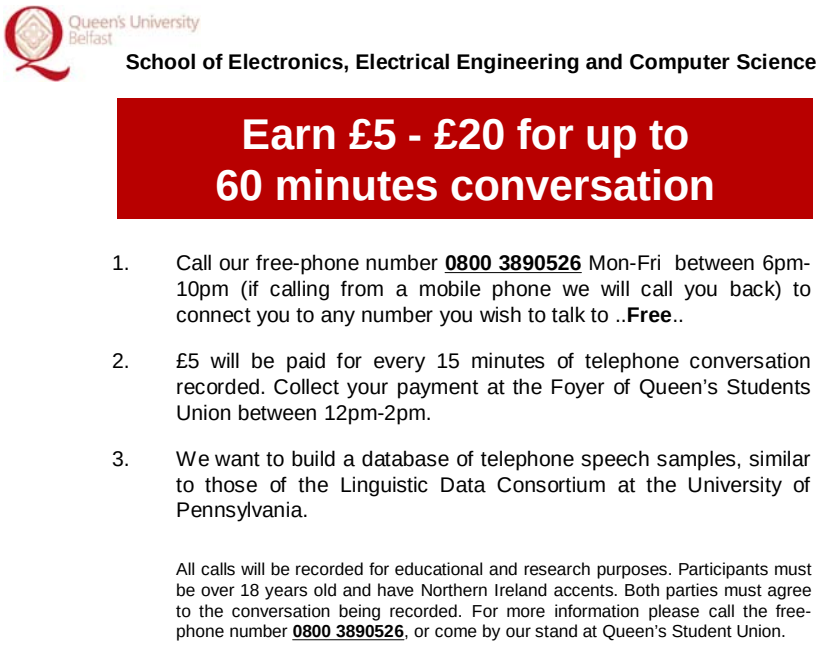

Figure 3. Example of one of our leaflets used to attract students

4.2 Payments

We explored different methods of paying the students. We found that following the method sometimes used by the LDC and paying by a cheque from the university was administratively too expensive. We also looked at the possibility of issuing a token, in the form of a small card, for each £5 payment. This could be redeemed in the main bar of the student union. The union would have agreed to this since students were then likely to spend some of their money in the union. The union would have invoiced us for the cards it collected every month. However, we finally decided to use cash after we unexpectedly got the agreement of the finance office of the

university to issue us with working cash payments in sums of £500, under the signature and the responsibility of the project leader, Professor Ming. The students collected their payments from the desk we set up at lunch time each day in the entrance hall of the student union, usually on their lunch break on the day after making their recording. To get their money, students had first to give their names and telephone numbers to identify themselves. These were checked against our paper records of the calls that had taken place. The students then signed a receipt for the money and filled out a short consent form giving their age, sex, and postal code of their home town and the same data for the person they were calling. Unless they filled out this form, they were not paid and the recording not used. However, they always did. The receipts were collected and returned to the finance office of the university where they were checked against the cash payments.

Being paid in cash at lunch time suited the students. It also suited the student union, which allowed us to set up the desks in their entrance hall, because the students went on to spend some of their money in the union. Apart from the fact that we had to take care of the money, it was easy to administer. Occasionally we ran out of money at the lunch time desk and had to take money temporarily from our own wallets. However, at the end of the project, all money was accounted for although 19 students did not collect the payments due for their recordings.

4.3 Sample conversations

Short samples of two typical conversations chosen at random are shown in Figure 4. Note that we include in the transcriptions the numerous and important disfluencies in spontaneous speech: hesitations recorded as a line, some recorded as a pseudo-word, e.g., *ewh*, *eh*, *ah*, *hey*, etc. as in the samples below and others. Also we include comments on extraneous noises, e.g., [noise of door].

Note that the transcription of the spoken conversations into legible text was surprisingly time-consuming. Although a foot operated treadle was used by the listener to move backwards and forwards in the recording, it still took 4 or 5 hours or more to transcribe one hour of conversation.

Sample 1

- S2: There is our Rachel out Rachel has got her hair
- S1: What
- S2: Dyed dark brown
- S1: Rachel?
- S2: Rachel XXXXXXX [Surname deleted]
- S1: ewh
- S2: Dark brown
- S1: What did she do that for?
- S2: I don't know why she has done it for, maybe to get it from not being ginger but her hair was a nice colour
- S1: I know like an auburny dark
- S2: Aye
- S1: Really dark chestnutty colour wasn't it
- S2: They have got it all in eh a dark brown now
- S1: What age is she now anyway?
- S2: I think she is about 15
- S1: Is she?
- S2: 15 aye about what? 3rd year?

Sample 2

- S2: How's it going Peter?
- S1: Well Niall how's it going aye grand suffering a wee bit with my hay fever but ah fuck
- S2: Did you watch the Brazil match last night
- S1: Aye good match I I really don't think
- [Noise of door]
- S2: I don't think ah
- [Noise of door]
- S1: they are going to win you know they say Argentina hey are looking the best so far
- S2: Aye well the jolly swag man came to a sad end by a _____
- S1: What are you talking about
- S2: No it's just that that's what the one of the that's what Tom _____ said about it in the Irish News today
- S1: Aye
- S2: He says Australia were done in by two Brazillians called Adriano and Fred, Football is sometimes poetry often

Figure 4. Two samples of recorded conversations by students

4.4 Unacceptable recordings

The operator manning the phones in the evening occasionally checked the conversations by listening to them; also before the recordings were stored in our database, each recording was listened to, not all of it but at least part of it, to ensure that the recordings were genuine conversations. In a few cases students had not carried on a proper conversation but had deliberately talked nonsense or gibberish. These recordings were not included, and if caught in time, no payment was made.

One student was found to be speaking Hindi when he used our system to ring home to India. He got a free phone call, but was not paid. On another occasion, the staff on the desk making the payments at lunchtime noticed that some of the girls collecting their payments were very young. It turned out that one of the girls at a local school had heard of our project from her older sister, a student at the university, and the girl had earned herself some extra pocket money. She in turn had been telling her friends at school. Unfortunately for legal reasons these recordings had to be deleted; our leaflets stated that participants had to be aged 18 or over. We did not get our money back!

4.5 Database and corpus

The recordings of the conversations were stored in a corpus that is accessed through a database. The database is simple, consisting of a single table of records with the following fields taken from the consent forms.

FIELDS IN DATABASE
ID 1 (participant 1 in the conversation)
Sex 1
Age 1
Post code 1
Major town or county 1
ID 2
Sex 2
Age 2
Post code 2
Major town or county 2

Duration of the conversation in minutes
Link to the corresponding speech file

Participant 1 was the principal person initiating the recording. The ID for this person was a number assigned automatically that acted as the key to the record. If the age was not known or not given, then it was usually estimated by the operator. To maintain the confidentiality of the corpus, only the first 3 characters of the post code were recorded to give the broad region of the accent of the person but not the address. The names and telephone numbers were only kept on paper to make the payments and were then discarded. The county or major town near where the person lived was used in case the postcode was not given or was incorrect.

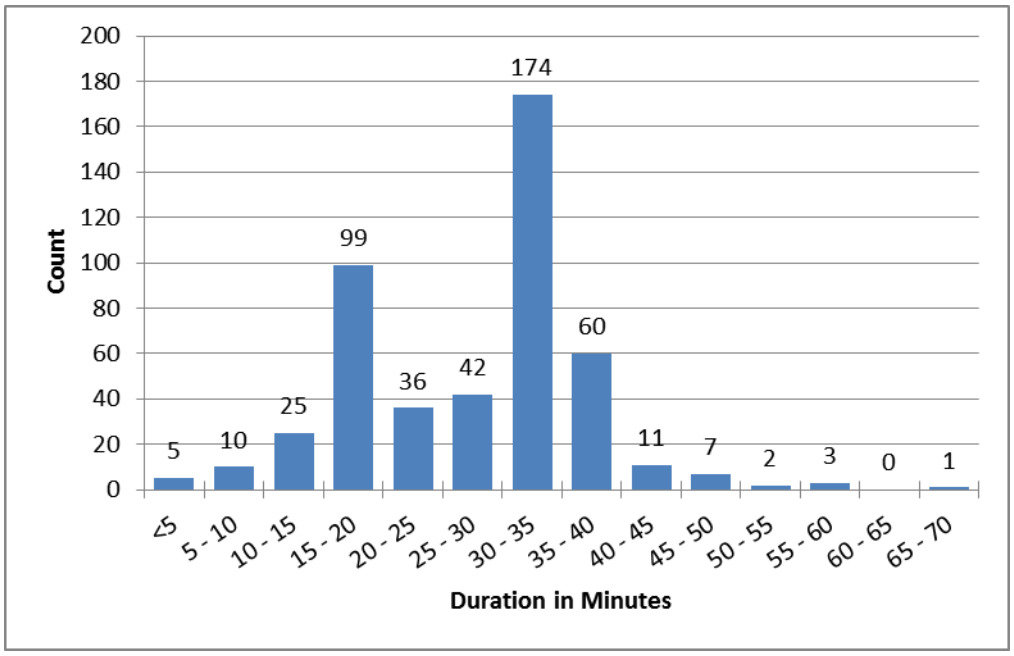

Figure 5. Distribution of the durations of the 475 recordings in the corpus

There are 475 recordings in the corpus making up 218 hours of conversations by 653 participants, with about 2.7 million token words (about 3% disfluencies) and 55,000 word types. The distribution of recording times is given in Figure 5. It is evident that most of the recordings were close to the requested 15 or 30 minutes in length. However, often it was difficult to stop students, and we only insisted when there was a queue of students waiting to record.

The age and sex distributions of the participants are given in Figure 6. The gender makeup of the actual recordings was as follows:

- 177 recordings with male and female speakers
- 178 recordings with all female speakers
- 120 recording with all male speakers

Near the end of the project, we discovered that more of our recordings were from female students than from male students, and to produce a better balance, we had to restrict our final recordings to male students only. If we were to repeat the exercise, we would correct this imbalance earlier.

As evident from the first peak in Figure 6, most of the speakers were aged between 18 and 22 years, corresponding to undergraduate students. The second peak between 25 and 29 were mainly graduate students and research assistants. The third peak arose because students often called home and the second speaker was therefore a parent or other relative, giving a small peak at about aged 50. Evidently a few grandparents took part in the conversations also.

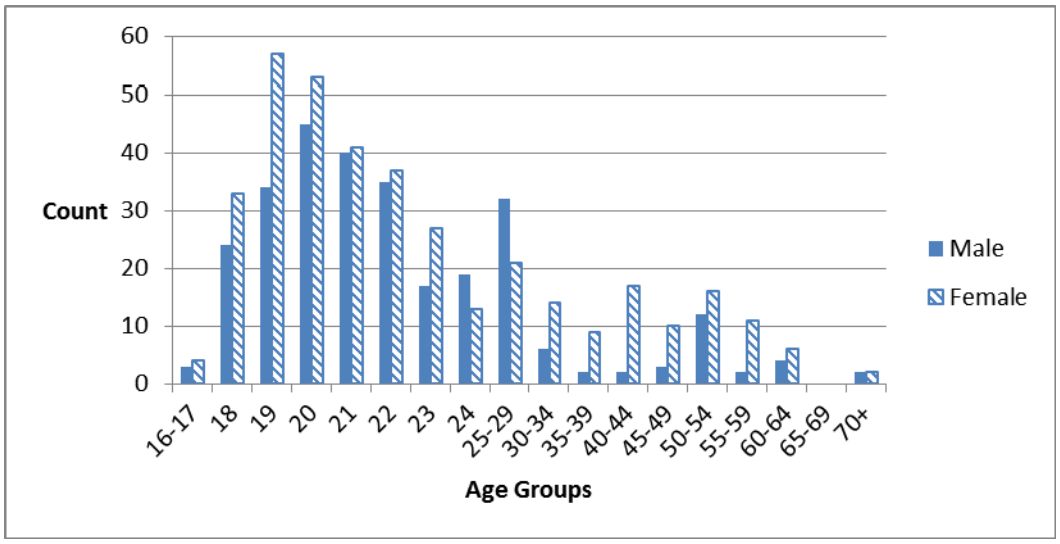

Figure 6. Distribution of ages and sexes of participants

The database for the corpus and some sample recordings with their transcriptions are given on our web site at:

http://www.qub.ac.uk/speech/index.htm

This web site also will give access to the entire corpus of recordings when ready for release by the beginning of 2012 along with conditions of use of the data. As mentioned earlier in Sections 3.2 and 4.1 (Figure 3), a restriction agreed to by each participant was that the data be used for "educational and research purposes". There will be a delay of 4 years after the last of the recordings, making it less likely that the participants can be identified.

5 CONCLUSION

Although the students knew they were being recorded and agreed to being recorded, they were in their own natural environment and were doing what they often do in the evening; they called one another on their cell phones. Therefore, it appears that they quickly forgot they were being recorded and, we think, provided us with a database of almost natural spontaneous conversations. We plan to make this corpus available through the Internet, accessed from our web site at the above address, by the beginning of 2012.

The primary purpose of the corpus is to record the continuous sounds of the speech itself including the diction (words being used), the phrases, the prosody, the intonation, the grammar (or lack of it), and the accents. These are all of interest for speech and language research. We also believe that, in time, the subject matter of the conversations will be of interest to social scientists and historians.

Throughout this project we were dependent on the voluntary participation of students, but we found that they did not always behave in the way we expected or would have preferred. Difficulties arose because of an unwillingness or reluctance of students to take part (Sections 2.1, 2.2, 3.1, and 4.1), some unexpected behavior (Sections 4.4), a male-female imbalance in the number of participants (Section 4.5), and a need for safety and security (Section 2.2). These unexpected problems taught us that when data are concerned with human behavior, they are sometimes more difficult to measure than the physical data with which we are more familiar. Nevertheless, we are thankful to all of the Irish students who gave their time to create this unique large corpus at the beginning of this 21st century. At the time of writing, our TitaniQ corpus is the largest freely available spontaneous speech corpus of English conversations.

As we listen to the recordings, it is clear that they are understandable by humans. The challenge now is to make them understood by our computer systems, a goal still a long way from the 65% accuracy possible at present.

6 REFERENCES

ACMSIGIR workshop (2007) *Searching Spontaneous Conversational Speech*, Amsterdam, The Netherlands. Retrieved from the World Wide Web, December 8, 2011: http://hmi.ewi.utwente.nl/spraakgroep/documents/sscs2007/sigir.pdf

Cieri, C. & Lieberman, M. (2000) Issues in Corpus Creation and Distribution: The Evolution of the Linguistic Data Consortium. *Proc. 2nd Int. Lang. Resources and Evaluation Conf.*, Athens: 1-8. Retrieved from the World Wide Web, December 8, 2011: http://www.LDC.upenn.edu/

Furui, S. (2003) Recent Progress in Spontaneous Speech Recognition and Understanding. Proc. *IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo: 253-258.

Furui, S. (2005) Recent Progress in Computer-based Spontaneous Speech Recognition. *IEICE Trans. Inform. & Systems* E88-D (3): 366-375.

Jelinek, F. (1985) The Development of a Discrete Dictation Recognizer. *Proc. IEEE* 73(11): 1616-1624.

Jelinek, F., Mercer, R. L., & Bahl, L. R., (1983) A Maximum Likelihood Approach to Pattern Analysis and Machine Intelligence for Continuous Speech Recognition. *IEEE Transactions* 5: 179-190.

Kuhn R. & De Mori R (1990) A Cache-Based Natural Language Model for Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12(6): 570-583.

Owens, M., O'Boyle, P., McMahon, J., Ming, J., & Smith, F. J. (1997) A Comparison of Human and Statistical Language Model Performance using Missing Word Tests. *Lang. Speech* 40(4): 377-389.

Young, S. J. (1996) Large Vocabulary Continuous Speech Recognition. *IEEE Signal Proc. Mag.* 13(5): 45-57.

(Article history: Received 25 February 2011, Accepted 3 December 2011, Available online 24 December 2011)