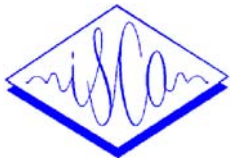

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

BUILDING SPEECH DATABASES FOR CELLULAR NETWORKS

Eric Sanders[1], *Henk van den Heuvel*[1], *Khalid Choukri*[2]
[1]SPEX, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
[2]ELRA, 55 Rue Brillat-Savarin, 75013 Paris, France
eric@spex.nl, H.v.d.Heuvel@let.kun.nl, choukri@elda.fr

ABSTRACT

The number of telephone applications that use automatic speech recognition is increasing fast. At the same time the use of mobile telephones is rising at high speed. This causes a need for databases with speech recorded over the cellular network. When creating a mobile speech database a number of problems show up that are not an issue when creating a speech database of fixed network recordings. These problems have to do with different recording environments, different networks and handsets, speaker recruitment and distribution, and the transcription. In this paper, the problems are explained, a couple of possible solutions are given and our experiences with these solutions in our contributions to the creation of mobile speech databases are presented. Besides, ELRA's position in the distribution of mobile speech databases is outlined.

1. INTRODUCTION

For as long as automatic speech recognition (ASR) research has been carried out, spoken language resources (SLR) have been needed for research and development. Both the use of cellular telephones and the development and use of ASR have grown rapidly over the last couple of years. The growing use of mobile telephony introduces an extra demand for ASR because mobile telephones are used in places where hands and eyes are needed for other things than operating the telephone (e.g. driving a car) and ASR can be used to control the telephone or other systems (like a radio in the car). Due to the specific conditions (high background noise, medium band coding and the impact of radio transmission problems) special databases are needed to support the development of ASR technology for use in cellular networks.

When building speech databases, many decisions have to be made and various problems have to be solved: the list of items to record has to be specified (this depends heavily on the purpose of the database), a recording platform has to be built, speakers have to be recruited, the recordings need to be transcribed, the database structure has to be decided, etc. See [2] for extensive information.

In the SpeechDat(II) project [3,6], a consortium was founded to handle these problems together. In this project 20 fixed, 5 mobile and 3 speaker verification databases were recorded and all languages in the European Union plus some dialectal variants were included. The databases were designed such that they contain a wide variety of items which are suited to train ASR systems for various applications.

SPEX [7] has been involved in building databases for both fixed and cellular networks. In the SpeechDat consortium SPEX validated all the speech databases, while ELRA [8] is in charge of the distribution aspects.

The creation of a mobile network database (MDB) introduces some problems that are far from trivial. In this paper we give an inventory of the difficulties involved in building a MDB in contrast with a fixed network database (FDB), some possible solutions to these problems and we propose 'best practice guidelines' based on our experiences. The paper is structured as follows: In the next section we pay attention to problems regarding different recording environments, different networks and handsets, speaker recruitment and distribution, and the transcription. Section 3 deals with ELRA's activities concerning the distribution of speech databases. In section 4 follow some concluding remarks.

2. MOBILE DATABASE RECORDINGS

2.1 Environments

One of the big differences between fixed and cellular telephony is the mobility of the caller. The speaker can call from many different environments. This means that all kinds of background noise can be expected in the recordings. Background noise is important both for training (noise models) and testing ASR systems. Different kinds of background noise need therefore be recorded and annotated.

Attention must be paid to the environments which should be in the database. Because the number of possible environments is huge, it is necessary to group environments with the same (expected) typical background noise. A compromise needs to be found for a set of a manageable number of classes that are

separate enough without much variance within a class. Furthermore, the classes should be likely to be called from in real situations and they should be safe to call from (e.g. one should not ask someone who is driving a car to read a number of sentences from paper).

In the SpeechDat(II) project four environment classes were used [6]: 1) a quiet environment (home, office), 2) a (crowded) public place (pub, train station), 3) alongside a busy street and 4) from a moving vehicle (car, bus, train), only passengers. The factual use of the database in the development of ASR systems has to prove whether this is a good set of classes.

The distribution of environments should be clearly defined. Either a minimum number of recordings from each environment should be set or the environments should be spread evenly. One way to get the distribution as required is to ask the callers to call from all the environments. This is only possible when the set of environments to record from is small. In the Dutch SpeechDat MDB we asked the participants to call from four different environments, but the readiness to call from all four was pretty low. After reminding many people to complete their recordings only half of the participants with at least one recording completed all four calls.

If it is not necessary that a participant calls from all environments (although in the case of a Speaker Verification Database it can be necessary) then a better option seems to be to ask each participant to call from only one (maybe specified) environment. When there are ample recordings from one environment, new participants are asked to record from the other environments until there are sufficient recordings from all environments.

A third possibility is to ask the participants to call from any of the environments and to keep recording until there are enough recordings from all environments. This will probably mean a lot of oversampling.

One should bear in mind that some environments are more difficult to call from than others and that it is therefore harder to meet the demands of a minimum number of speakers for these environments. To solve this problem one could give the participants a higher reward if they call from a difficult environment.

In any case it should be checked from which environment the recording was actually made. It is therefore necessary to include a question to the caller from which environment (s)he calls. It can then be decided to leave the classification of the environments to the caller in which case the alternative classes should be given. It can also be decided to leave the classification to the postprocessor. If, for example, the caller says (s)he is calling from a postoffice, the postprocessor could classify the environment as public place. We have used both methods and they seem both equally useful.

2.2 Networks

Mobile networks differ more from each other than fixed networks. There are analogue and digital mobile networks, although the former will probably disappear in a few years' time. In Europe GSM seems to become the only mobile network, but in the whole world a couple of other standards exist. Furthermore, most countries have a number of providers with their own networks of different quality. For example, a small country like the Netherlands with 15 million inhabitants currently has 5 mobile network providers.

It is desirable to know from what kind of network and provider a call is coming from. In some countries it is possible to derive the network (provider) from the telephone number. It is then possible to detect the network if Calling Line Identification (CLI) is available. If this is not possible, the caller should be asked from which network (s)he is calling.

When demands are made on the distribution of the networks in the database (in SpeechDat(II) at least 90% had to be GSM calls) a few strategies are possible to satisfy this requirement. If it is possible to infer the network from the telephone number, speakers should be recruited according to their telephone number. Another way is to have a mobile phone with a subscription to all the networks and have the speakers call with this phone over one of the required networks.

A problem that occurs often in cellular telephony is losing the connection. Especially when going from one cell to another (thus when moving) the connection gets lost easily. If a participant is recording 50 items and the connection is lost at the 40th item, (s)he is probably not very motivated to do the whole recording again. A useful solution to this problem is to give the caller the possibility to resume the recording at the spot where the connection was lost. This should be done within a certain time interval, otherwise the conditions of the last part of the call could vary too much from the first part of the recordings. We implement half an hour as maximum interval, which seems a reasonable time frame.

2.3 Handsets

There are many different handsets. For some reasons it seems less sensible to put much effort in getting these distributions (right). One is that the life time of a handset is so short that the information of which handset is used is almost out of date when the database is ready. Secondly there are so many handsets

available that there is no end in recording all the different handsets and new ones hit the market at great speed. Finally, the variance in network quality is assumed to be so much larger than the variance in handset quality at this moment that it does not really matter which handset is used. Recently, some networks transmit the GSM signal at enhanced full rate (EFR) instead of full rate (FR), which gives a big improvement in quality, but for which a special handset is needed. Information about which rate is used to transmit the signal is interesting to know, but hard to retrieve since the network will switch from EFR to FR if it is overloaded.

2.4 Speaker recruitment and distribution

Whereas recruiting speakers for a FDB is already difficult, for a MDB it is even harder. Almost everybody owns a fixed network telephone, but not everybody owns a mobile telephone. Besides, the distribution of owners of a mobile telephone is not evenly spread, e.g. a mobile phone is much more popular with young men than with older women. The number of mobile telephone owners and their distribution is country-dependent, though. Experience has taught that it is most effective to use as many methods as possible to recruit speakers. An advertisment could be placed in a periodical with many mobile phone owning readers. For general speaker recruitment strategies see [4].

The big advantage of mobile telephony, i.e. that the phone is portable and can be taken everywhere, should be exploited. E.g. by asking participants to have friends or family do the recordings with their telephone and give them a reward for this or by hiring someone to take to the streets with a mobile phone asking people to do the recordings.

It is important to decide how the distributions for age, sex and accent for a balanced database should be and to get it right. This problem does not differ much from FDB recordings except that the distribution of mobile phone owners in the real world is not evenly spread. Our experience is that more men have mobile phones than women, more young people than old people and more people living in some parts of the country than in the other parts of the country. The easiest way to get the distribution right seems to be to start recording everybody and after a while to only recruit people from a category that is not yet sufficiently present in the database and, if necessary, give extra bonus rewards for calls in the rare categories.

2.5 Transcription

Recordings made over the mobile network differ in such a degree from recordings made over the fixed network that extra attention should be paid to the transcription of the recordings. GSM introduces quite some distortions, like buzzes, fade outs and drop outs. These can be annotated by marking each word that is affected by it. In SpeechDat a special marker (the ampersand) was attached to each word that was affected by one of these events.

Because speech is recorded in so many different environments, many kinds of background noise can be expected. A good way to treat these is to mark only those background noises that are not expected in the environment the recording was made in. E.g. it is not necessary to annotate traffic noise on recordings made from the street side, since that is exactly the expected noise. There is a grey area of background noises that are only partly expected, like radio music in a moving vehicle. Unexpected background noise should be annotated as is normally done.

3. ELRA

The distribution of the SpeechDat MDBs will be carried out by the European Language Resources Association (ELRA) as stated in the Technical annex of the SpeechDat(II) project. The distribution conditions and terms will be negotiated between ELRA and each database owner.

ELRA was founded in Luxembourg in February 1995 with the goal of promoting the creation, validation and distribution of language resources (LRs), with a preliminary mission to contribute to the development of the emerging market of language engineering. A crucial task for the association is to identify and collect existing resources and to spread this information to potential users in Europe and beyond. In order to do so, ELRA has to address legal, technical, logistic, and commercial issues [1].

As of March 1999, the ELRA catalogue lists 92 speech resources, out of which 11 are SpeechDat databases (mainly FDBs) and 4 are SpeechDat-like databases.

For the MDB resources, the legal issues concern the relationship between the producer and ELRA, and between ELRA and the users. The relationships between ELRA and the provider of the MDB, or between ELRA and the users of the MDB will be handled through appropriate distribution or usage licenses, based on the generic ones available on ELRA's Web site.

The pricing policy is another crucial issue. ELRA has always tried to negotiate low prices with the providers in order to boost the deployment of speech technologies for as many languages as possible but also to encourage research activities. Even in a new and emerging market, this has been successfully achieved as illustrated by the 47 FDB distributed so far, covering 8 different languages. Basically, SLR prices are correlated to the production costs and the market value of the data. The market of MDBs is similar, in its players, to those of FDBs. The speech-based services are more critical in mobile telephony than in other areas, e.g. users can not easily use DTMF services and may instead strongly require speech recognition facilities. Global mobile telephony providers are forced to plan the introduction of speech technologies in their services. This may have an impact on the pricing policy which is presently difficult to predict. Nevertheless, ELRA will continue to argue for low prices for the use of the data for R&D purposes. ELRA will also continue to offer its members a substantial discount on the public price.

ELRA will also continue to stimulate the production of new resources as it will continue to issue calls for the production and the packaging of Language Resources whenever its finances permit it to do so. This also holds for MDBs, since they considered a valuable supplement to the present market.

4. CONCLUSION

Creating a MDB introduces some problems not relevant to FDB. In this paper we summarised the most relevant differences and issued some solutions and best practice guidelines to cope with them. All these problems must be considered when building a MDB and all decisions taken must be motivated and documented. Undoubtedly, further problems will present themselves now that the mobile database will be used on a large scale in the development of ASR systems. That may pose new demands for the creation of cellular databases. It must also be mentioned that some specific problems with MDB production will diminish with time. The quality of the network will increase and the number of users will grow, so that speaker recruitement will become an easier task.

5. REFERENCES

[1] K. Choukri. New developments within the European Language Association. *Proc. Eurospeech 1999* (Budapest, Hungary).
[2] D. Gibbon, R. Moore, R. Winski (eds). Handbook of Standards and Resources for Spoken Language Systems. Mouton de Gruyter, 1997.
[3] H. Hoege, C. Draxler, H. van den Heuvel, F.T. Johansen, E. Sanders, H. Tropf, SpeechDat Multilingual Speech Databases for Teleservices: Across the Finish Line. *Proc. Eurospeech 1999* (Budapest, Hungary).
[4] B. Lindberg, R. Comeyne, C. Draxler, F. Senia. Speaker Recruitment Methods and Speaker Coverage – Experiences from a Large Multilingual Speech Database Collection. *Proc. ICSLP 1998* (Sydney, Australia), pp. 2731-2734.
[5] J. vd Velde, D. Langmann, M. Palewski. Specification of Speech Data Collection over Mobile Telephone Networks. *SpeechDat Technical Report SD1.1.2/1.2.2* 1996.
[6] SpeechDat Family: http://www.speechdat.org/
[7] SPEX: http://lands.let.kun.nl/spex/
[8] ELRA: http://www.icp.grenet.fr/ELRA/