The design, collection and annotation of speech databases in South Africa

Febe de Wet[†]*, Pippa Louw*[†] *& Thomas Niesler*[‡]

[†]Centre for Language and Speech Technology (SU-CLaST)
[‡]Department of Electrical and Electronic Engineering
Stellenbosch University, South Africa
{fdw,phlouw,trn}@sun.ac.za

Abstract

This paper presents a number of practical guidelines on the design, collection and annotation of South African speech databases. Most of the guidelines are based on personal experience gained during previous data collection exercises. The issues that are addressed in the paper include: the aim of data collection, the design of prompting material, speaker recruitment, recording equipment, as well as the recording, editing and annotation of the speech data.

1. Introduction

The aim of this paper is to systematically describe some of our past experience with the design, collection and annotation of speech databases in South Africa. Some of the issues that we will raise may sound too obvious to be worth mentioning. However, past experience has shown that neglecting the seemingly most obvious factors can have a disproportionally large impact on the quality of the final database. While the practical issues of database compilation cannot be viewed as cutting-edge research, they can consume a large portion of a project's resources, and seriously affect the standard of the overall result. We will often illustrate our statements with anecdotal evidence, which is based either on personal experience or on interviews with annotators.

Among the many points that have been raised during our discussions, the most frequently recurring piece of advice has been to incorporate a deliberate and well executed preliminary trial run into any data collection effort. This will allow the peculiarities of the project to be assessed in the light of hard evidence, and will greatly lessen the need for difficult, time consuming and expensive later remedial action. In particular, if more than one language is concerned, a pilot collection should be undertaken for each individually because cultural differences may make it impossible to draw general conclusions. For example, in many of the more extensive data collection efforts that have been undertaken in Europe (e.g. TC-STAR [1, 2] and the SpeechDat family of databases [3]), all partners were required to submit a 10-speaker prototype database for validation before they were allowed to continue with the collection of the complete database. This procedure ensured that systematic errors could be identified and rectified at an early stage of the speech collection process [4].

The following sections of this paper will deal with issues including the role of the overall aim of the data collection effort, the design of appropriate prompting material, the recruitment of speakers, the equipment used for recording, and the process of editing and annotating the data. Database design and technical specification will not be addressed here since these are highly project-specific. Information regarding the database structure, file formats and similar technical detail used in past projects can for example be found in the validation documents that accompany the European Language Resource Association's spoken language resource catalogue [5].

2. The aim of data collection

Before starting an extensive data collection process, the intended use of the speech material should be clearly specified because it determines the quantity, style and type of data that needs to be collected as well as the level of detail at which it should be transcribed and annotated. Factors that should be considered may include (but are not limited to):

1. To what extent must the triphones of a specific language be covered?
2. Is read speech or spontaneous speech required?
3. Is studio-quality (wideband) or telephone speech required?
4. If telephone speech is required, should the calls originate from cellular or fixed line networks, or both?
5. Should all speech be gathered through the same channel (microphone or telephone handset) or from a variety?
6. Should each speaker be prompted for the same material, or should speaker-specific prompts be used?
7. What age groups, level of language proficiency and mother tongues or accents should the speakers have?

The nature of the research or development that will be carried out with the data will determine the answers to the above and related questions. For example, if telephone-based dialogue systems are to be developed, telephone data from a variety of speakers and handsets must be gathered. However, if pronunciation variation among a defined group of mother-tongue speakers is to be researched, phonetically-motivated prompts should be recorded in a studio via the same microphone to avoid interfering channel effects.

3. Prompting material

When prepared speech is collected, subjects will generally be asked to read from a collection of specially-compiled text termed the **prompting material**. Factors which may be of importance in the design of this material include:

1. Are isolated words, sentences, or entire passages of speech desired?

2. Should the material be phonetically balanced or representative?
3. The level of difficulty of the prompts should be commensurate with the expected level of literacy of the intended subjects.

The design of prompting material should also take the data gathering strategy into account. For example, more complex scripts can be used for data collection in supervised, studio recordings, while simple texts that are easy to understand should be used for participants who are required to make unsupervised telephone calls. We once developed what we considered to be a very straight-forward prompt sheet, only to discover that a large number of callers were reading both the instructions as well as their answers [6]. The use of long sentences with large phonetic coverage resulted in utterances that were heavily corrupted by hesitations, mispronunciations, repairs and restarts from all but the most skilled mother-tongue speakers. Even in supervised recording situations we have found it more effective to use a larger number of shorter prompts rather than fewer, longer prompts.

The nature of the prompting material should be focussed on the overall aims of the project, as indicated in Section 2. For example, in a project investigating variations among South African English accents, an attempt was made to ensure high phonetic coverage suitable for acoustic model adaptation, as well as to gather data that reveals something about the subjects' English proficiency. It turned out to be rather difficult to kill both birds with one stone. The phonetically rich text was reproduced very poorly by the less-proficient subjects, casting doubt over their suitability for acoustic modelling. In retrospect, it would have been more effective to have two separate data collection exercises, each of which could then have been more focussed. Previous attempts to collect all-encompassing databases that are suitable for both phonetic research as well as the development of telephone-based spoken dialogue systems have encountered similar difficulties.

Although mundane and time-consuming, thorough proofreading of all prompting material is a crucial part of any data gathering exercise. If the prompting material is not error free, the rest of the process will invariably suffer. This is ultimately very costly, because the recording, editing and annotatation of the data requires a vast effort. During a Xhosa data collection exercise, for example, a unique prompt sheet was automatically generated for each participant. However, a certain number of Zulu words were erroneously included in the Xhosa prompts, and this fault was not immediately detected. The result was that several subjects began to object to the prompts while making their call. This in many cases rendered the data captured from these speakers unusable. The significance of prompt sheet design and verification is also demonstrated by the recent extension of the standard European validation procedure to include prompt sheet validation as a preliminary step, which precedes the validation of the 10-speaker prototype database referred to in Section 1 [4].

When read speech is to be recorded, care should be taken to anticipate differences in reading proficiency. Factors such as the length of the sentences, the familiarity of the vocabulary, as well as the complexity of sentence construction play a role in this respect. Furthermore, people generally find it harder to read material that they are not familiar with. In the South African English accents project mentioned before, participants were asked to read extracts from the Goldilocks fairy tale, which has often been used in accent studies by other researchers [7, 8]. However, the passage, which is well-known to most Europeans and Americans, is often not familiar to South Africans whose mother tongue is not English. Since many participants in the study were L2 and L3 users of English, they had considerable difficulty reading the passage.

When prompting for spontaneous speech, reliance on the availability of external information should be avoided. For example, if the time of day is prompted for, the subject may not be in possession of a watch, or if the weather is enquired about, the subject may not be close to a window.

In addition to being culturally sensitive, prompting material should also not be offensive or humorous. Fits of laughter can obscure speech data to such an extent that the recording must be discarded. Scripts should not contain foreign words or proper names. For example, some of the prompt sheets that were used in previous projects contained phonetically rich sentences from the TIMIT and SCRIBE corpora. Some of these sentences contain names and words that are specific to American and British English. Transcribers have found overwhelmingly that people often do not know how to pronounce these items properly, or forget how to when faced with the stressful recording environment. The utterances in which these words occur consequently contain many disfluencies.

The physical presentation of the prompting material is another issue that deserves attention. Especially in studio recordings where sensitive microphones are used, speakers should be made aware of the fact that fidgeting with paper prompt sheets (or with any other object, like clothing or a cell phone) can interfere with the recording. To remove this source of interference, we usually have a staff member present in the recording room with the speakers to handle all printed material on their behalf. If it is possible to set up a noiseless computer screen in the studio, prompts could also be displayed electronically. This strategy is especially efficient for recording word lists. No matter how clearly speakers are instructed to read words slowly and individually from a list, they invariably begin to read faster and forget to articulate individual words properly. In the worst cases, long lists of words have caused speakers to move their heads substantially relative to the microphone, creating an amplitude modulation-like pattern in the recordings.

Finally, to avoid fatigue, the overall length of data collection from each speaker should not be too long. Our experience is that each subject should not be required to read or speak for more than approximately 20 minutes.

4. Speaker recruitment

A speech database ultimately contains recorded material gathered from a certain set of **speakers**. The task of locating these speakers and then motivating them to take part in the recordings is normally the task of the speaker **recruiters**. The recruiters are often simply the members of the research team, but this may not be the most efficient arrangement. The relationship between the speakers, the recruiters and the research team members is critical to the efficiency and ultimate success of the data gathering exercise.

Before recruitment starts, the speaker population must be well defined. This ensures that recruiters do not target speakers who are inappropriate for the project. Furthermore, different recruiting methods may be more effective for different types of speakers. For example, a telephone-based database project

found a very high rejected call rate amongst elderly callers. If such speakers are a deliberate part of the target population, special effort must be made to improve on the success rate of their calls.

Recruiting suitable speakers has been a major bottle-neck in other projects [9]. Several approaches to speaker recruitment exist, for example:

1. A large population of potential speakers can be contacted by unsolicited email. When speakers are not rewarded financially for their participation, however, response rates have been found to be extremely poor. This is also true when an uncertain reward such as a prize or entry into a draw is offered. On the other hand, we have found that subjects who are rewarded financially for their efforts are considerably more cooperative and motivated. However, it is important to reward only those subjects who complete their recordings successfully. Furthermore, for telephone-based data gathering it is difficult to prevent the same speaker from placing multiple calls and to detect fraudulent claims of multiple rewards.
2. A marketing company can be employed to recruit speakers. Marketing companies have a much better knowledge of the target population than a typical speech researcher in charge of the project. However, this approach may be too costly, and feasible only for telephone-based data.
3. A "snowball" approach can be used fairly successfully. Here each subject is rewarded not only as a speaker for his or her own contribution, but also as recruiter for each additional subject recruited. In this case, it is particularly important that recruitment rewards apply only to successful recordings, and not simply for the number of additional subjects recruited. This will motivate recruiters to supervise calls, and to reduce the incidence of fraudulent calls in which a single subject attempts multiple recordings, each time disguising his or her voice.
4. A defined small group of people are appointed as recruiters. Since the recruiters are known to the research team (this is not necessarily the case in the previous technique), they can be more carefully selected.

The care with which speaker recruitment is done impacts heavily on the average quality of the recordings that are made, and on the number of aborted or excessively corrupted recordings. For this reason recruiters should ideally be thoroughly screened, and well-trained. They should be given clear instructions and should ideally be professionals who are paid well enough to take their job seriously. Before recruiters start their field work, they should participate as subjects in a pilot data gathering exercise. This will ensure that they have first-hand experience of the process, and that they understand what will be expected of the subjects whom they will be recruiting and supervising.

In the past, we have relied on academic and student volunteers as recruiters more than once. However, we found many resulting recordings to be useless because the recruiters did not take their task seriously enough and did not follow the instructions they were given. For example, in one case the recruiters were specifically instructed to ensure that all telephone calls were to be made from a quiet environment. Despite this, many calls were made from noisy supermarkets or from homes with the television or radio blaring in the background. In other cases recruiters repeatedly interrupted the callers by giving them instructions during the call. Some even went so far as to persuade illiterate people to participate, and continuously whispered the intended reply to the speakers during the call. In addition to making life difficult for the transcribers, this "technique" often resulted in confluent speech which rendered the data useless for the project. Ideally such recruiters should be identified and dismissed from the project team as quickly as possible.

Recruiters must verify the identity and mother tongue of all participants. Ideally, each participant should complete a personal questionnaire to determine how well the desired and the actual speaker populations match. During the collection of Zulu data, for example, we found that some of the subjects were not Zulu, but in fact Swazi. The subjects attempted to simulate Zulu speech, but their speech gradually reverted to Swati during the recording, as their attention was distracted by the task of reading. This "mistake" was discovered by a Zulu mother tongue transcriber, long after the recordings had been made. In a studio situation, subjects can be interviewed, making such "accidental translation" incidents unlikely. For telephone calls, the recruiters can play the role of monitors.

Although recruiter and speaker rewards can be a significant expense, high quality data from the appropriate speaker population is a very valuable resource. Furthermore, by reducing the incidence of dud and low-quality recordings, the subsequent annotation is simplified. During the African Speech Technology project, it was for example found that 41% of the almost 6000 calls that were recorded were later classified as empty or unusable [6].

5. Equipment

Before any audio data can be collected, a means of capturing and storing the speech is required. Since applications in speech, speaker, language and dialect recognition require the speech in digital format, analogue recording platforms should be avoided. Instead, computer based digital recording platforms are cost effective and convenient. For the capture of telephone speech, a telephone line interface will be required for the computer to accept telephone calls. For wideband speech, an audio interface that is external to the computer is recommended for improved signal-to-noise ratios, since the computer's integrated sound interface is usually contaminated by electrical noise. The computer should ideally be placed in a separate room from the microphone, to minimise fan-generated noise. For the same reason, the room containing the microphone should not have air-conditioning or be subjected to other background noise sources. To minimise reverberations, the recording room should ideally not have too many hard wall and floor surfaces. Wideband speech should be recorded as raw 16-bit (or 24-bit) audio. Lossy compression such as MP3 should be avoided since their effect on automatic speech recognition and other systems has not been well established.

The recording platform should be well tested before data gathering begins in earnest, to ensure that signal-to-noise ratios and other characteristics are within desired specifications.

6. Recording

Before each recording session or telephone call, subjects should be allowed to familiarise themselves with the prompting material and recording procedure. If recordings are taking place

in a studio, time should also be allowed for subjects to become familiar with the recording environment and equipment. A strategy should be devised in advance of how to deal with pronunciation errors, hesitations, restarts and other disfluencies, and subjects should be made aware of these arrangements before the recording starts. For example, if fluent and appropriately-pronounced material is required, subjects may be asked to restart at sentence boundaries if they make a mistake.

In a studio setting, the microphone is normally placed in a separate room from the remaining recording equipment and recording technician. These rooms are often adjacent but acoustically isolated to ensure good signal-to-noise ratios, although a soundproofed window is frequently present between the rooms. A subject in the recoding room may therefore experience it as an isolated and unfamiliar environment, leading to a certain degree of anxiety which can negatively affect the quality of the speech. We have found it helpful to have a member of the research team (to whom the subject has been introduced) to assist the subject in the recording room while his or her voice is being recorded. This also makes it possible to indicate to the subject when he or she should re-read certain prompts due to serious mispronunciations or disfluencies.

If a spoken dialogue system is used to gather telephone data, the first calls must be monitored during a trial run to make sure that the system is working as expected. It should be made clear to participants that they will be engaging with a machine, and that they therefore cannot ask for assistance during the call. We have frequently encountered recorded calls consisting of a cycle of misunderstandings on the part of both subject and machine. This normally begins with an incorrect response to the automated prompts, which in turn leads to further unexpected behavior from the dialogue system, and resultant inappropriate user reply. In such cases the data was invariably rendered useless by excessive stammering, by the speaker becoming too nervous or confused to speak properly, or by premature termination of the call due to frustration.

If possible, important information, such as the subject's name, identification code or data sheet number, should not be prompted for during the first few dialogue turns. Speakers often require a few practice turns to become accustomed to and comfortable with the system, and hence there is a risk that important data may be lost. It may even be advisable to include a short and explicit training section at the start of the dialogue.

Robust end point detection is essential when gathering data with a spoken dialogue system. End-point detection often fails in the presence of background noise. When the delay between the end of a user's response and the next system prompt is too long, subjects may assume the system is not functioning and terminate the call, or speak out of turn. We have even witnessed cases in which subjects became verbally abusive to a system that was apparently not responding. In fact, due to the continuing, impatient interjections of one specific user, the end-point detection was never able to locate the end of the user's utterance, causing the the dialogue system to become stuck in a particular state. To ensure a better proportion of successful calls, all calls should ideally be assisted.

Some method must be provided to reconcile the audio data with the data sheets, speaker questionnaires or prompting text. This can be achieved by ensuring that each subject speaks his or her name, identification code or data sheet number at some point during the recording. If possible, reconciliation between the audio data and completed questionnaires (speaker information) should take place before the data is transcribed. In one of our previous projects the raw data arrived in batches from a company at which the recordings were made. Some calls occurred in more than one batch and because different people were involved in the transcription process, these calls were transcribed by two different transcribers. The problem was only identified when the same transcriber eventually encountered the same call a second time. This sequence of events could have been prevented if the audio data had been checked against the speaker information before transcription started. Accurate call logging and unique call identification strategies can also be used to double check the link between the audio data and the speakers.

7. Editing and annotation

The tools that are used for editing and annotation should, as far as possible, be open source and available to anyone in the speech community who would like to access and use the data. Tools like Praat [10] and HTK [11] ensure continuity and shareability across sites and across platforms.

7.1. Editing

As a first step, the raw audio recordings must be processed into a form that allows them to be reconciled with the scripts and data sheets. This may involve the segmentation of the recordings into smaller units. It may also involve the removal of repetitions, repairs, hesitations and other disfluencies from the data, if these are undesirable. Recordings that are excessively corrupted by such artifacts, or by other factors such as background noise, can be discarded at this point. Hence this stage also serves as a first phase of quality control, and ensures that the material that is passed through for annotation is of a sufficiently high quality and in an appropriate format. It is useful to edit the data as soon after recording as possible, to allow for the re-recording of discarded material, as well as the timeous adjustment of any recording conventions that are spoiling the data.

7.2. Annotation

The type and degree of detail of annotations will be determined by the ultimate purpose of the corpus. For example, data gathered for acoustic modelling in spoken dialogue systems will usually be annotated only at the orthographic level, while data gathered to study accent variations will require phonetic transcription. Whatever the case may be, past experience teaches that the annotations should be kept as simple as possible, given the scope of the project. It is tempting to add to the detail of annotations for the sake of as yet unspecified "future research". For example, one may try to annotate the types of speaker and other noise that occur. Invariably, however, such "extra" detail leads to unforeseen additional complexity, and conspires to significantly retard the annotation process and increase its cost. What should be avoided above all is the "making-it-up-as-we-go-along" approach. This does not only lead to immense frustration on the part of the transcribers, but often has complicated consequences for the consistency and backward compatibility of the data.

Mark-up conventions that describe the data to a sufficient level of detail must be decided upon. Phenomena which may be annotated include (but are not limited to):

1. Full words (orthography).

2. Phonemes.
3. Sentence boundaries.
4. Phrase boundaries.
5. Word fragments.
6. Mispronunciations.
7. Proper names.
8. Speaker noises.
9. Other noises.
10. Code switching/mixing.
11. Speaker changes/turn taking.
12. Channel conditions.

However, no matter how carefully conventions are drawn up beforehand, unforeseen situations are likely to occur during the actual data gathering process. Therefore transcription must be part of the piloting exercise. Unforeseen changes to the mark-up conventions should be identified and incorporated into the annotation process as quickly as possible to ensure consistency. In an extensive project, this may be dealt with effectively by deciding in advance on a procedure for bug reporting and for version control.

Ideally, recording and transcription should take place concurrently. This is usually impractical because transcription is a slow and laborious process. However, if feedback from the transcribers can be incorporated into ongoing recordings, the quality of the data can often be considerably improved, and the need for remedial action minimised.

Transcribers should be trained by someone who understands the ultimate use of the data. This training should involve real transcription and extensive feedback to the transcriber. Transcription work should not begin in earnest before the proficiency of the transcriber in the mark-up conventions has been properly established. It has even been suggested that the transcribers should also be the recruiters. This will motivate these individuals to ensure a high quality in the recordings stage, since they will ultimately also be responsible for the transcription. Where this is not feasible, each recruiter should at least transcribe one recording to make him/her aware of the importance of thorough supervision of their subjects.

Transcribers should be paid well, but according to their output and not to the number of hours they work. Inefficient transcribers have consumed many transcribing hours without producing substantial output, and thereby depleted project resources. However, the transcribers' work should be submitted to quality checks at regular intervals. Quality control is especially important if transcription work is out-sourced to a third party. The whole process of transcription, quality control and payment should be streamlined to ensure that transcribers who deliver quality work on time are rewarded appropriately and timeously. A system can also be in place to track each transcriber's performance during the course of the project. Moreover, the researchers themselves should be involved in quality control, to ensure that the transcriptions fulfill the aims of the database. No-one who would ultimately like to use the database should be considered too senior to take part in transcription and quality assurance.

Finally, the annotation process can be aided greatly by online automatic format and spell-checking tools. These can be incorporated into the annotation software, and can reduce the effort required for error checking by improving the quality of the initial transcriptions.

8. Summary and conclusion

The compilation of a speech database is a major undertaking, whose scope in terms of time, manpower and money is often grossly underestimated. Past experience has taught us that the success of such a data gathering and annotation exercise depends on realistic goals, on careful planning and project management, and on good communication amongst all involved parties. The aim of the data collection exercise must be absolutely clear, and the design of the contents closely aligned to it. Care should be taken in the selection of both recruiters and subjects, and recording should take place according to specifications that are both clear and simple. Timely and (if possible) concurrent quality assessment will ease and quicken the painstaking process of annotation. Finally, all team members should have a clear understanding of what their tasks are and how these relate to the overall project aims.

9. Acknowledgements

We would like to thank (in alphabetical order) all those who shared their thoughts and ideas about their work on previous projects with us: Edward de Villiers, Ulrike Janke, Thembi Malaza, Fiona Marais, Luvuyo Martins, Michael Tait, Albert Visagie and Alison Wileman.

10. References

[1] H. van den Heuvel, K. Choukri, C. Gollan, A. Moreno, and D. Mostefa, "TC-STAR: new language resources for ASR and SLT purposes," in *Proceedings of LREC*, Genoa, Italy, 2006, pp. 2570–2573.

[2] http://www.tc star.org/, (accessed 27/10/2006).

[3] http://www.speechdat.com/, (accessed 27/10/2006).

[4] H. van den Heuvel, D. Iskra, E. Sanders, and F. de Vriend, "SLR validation: current trends and developments," in *Proceedings of LREC*, Lisbon, Portugal, 2004, pp. 571–574.

[5] http://www.elra.info/, (accessed 27/10/2006).

[6] P. H. Louw and M. Theunissen, "Annotating the AST speech databases: Practise makes perfect," 12th Biennial Conference of the African Language Association of Southern Africa, Stellenbosch, South Africa, 2003, (Presentation).

[7] C. G. Clopper and D. B. Pisoni, "The Nationwide speech project: a new corpus of American English dialects," *Speech Communication*, vol. 48, no. 6, pp. 633–644, 2006.

[8] P. Stockwell, *Sociolinguistics: A resource book for students*. London: Routledge, 2002.

[9] C. Draxler, H. van den Heuvel, and H. S. Tropf, "SpeechDat experiences in creating large multilingual speech databases for teleservices," in *Proceedings of LREC*, Granada, Spain, 1998, pp. 316–366.

[10] http://www.praat.org.

[11] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book, version 3.2.1*. Cambridge University Engineering Department, 2002.