Methodologies for Designing and Recording Speech Databases for Corpus Based Synthesis

Luís C. Oliveira, Sérgio Paulo, Luís Figueira, Carlos Mendes, Ana Nunes‡, Joaquim Godinho‡

L^2F INESC-ID, ‡INOV
Lisboa, Portugal
{luis.oliveira,sergio.paulo,luis.figueira,carlos.mendes}@l2f.inesc-id.pt, {ana.nunes,joaquim.godinho}@inov.pt

Abstract

In this paper we share our experience and describe the methodologies that we have used in designing and recording large speech databases for applications requiring speech synthesis. Given the growing demand for customized and domain specific voices for use in corpus based synthesis systems, we believe that good practices should be established for the creation of these databases which are a key factor in the quality of the resulting speech synthesizer. We will focus on the designing of the recording prompts, on the speaker selection procedure, on the recording setup and on the quality control of the resulting database. One of the major challenges was to assure the uniformity of the recordings during the 20 two hour recording sessions that each speaker had to perform, to produce a total of 13 hours of recorded speech for each of the four speakers. This work was conducted in the scope of the *Tecnovoz* project that brought together 4 speech research centers and 9 companies with the goal of integrating speech technologies in a wide range of applications.

1. Introduction

There are currently several speech synthesis systems with enough naturalness to be used in applications for a wide range of domains. The use of these systems has been restricted by the number of available voices and by the occurrence of artifacts when synthesizing less common words. Companies do not like to have an interactive voice response system (IVR) with the same voice as theirs competitors and most applications of speech synthesis require the use of domain specific words, like brand names or technical terms with unusual phonetic sequences. These restrictions are a consequence of the technology used in most speeech synthesizers that are based on the concatenation of variable length speech units taken from an inventory of recordings of a single speaker.

To assure the coverage of the most common sequences in a given language, the inventory must contain a considerable amount of speech recordings (from 3 to 10 hours or more) with carefully selected contents. These contents are designed to provide a good coverage of the phonetics and intonation of the selected language using analysis performed on available text corpora, mainly newspaper texts and books that do not always cover the specific requirements of certain applications such as speech-to-speech translation, medical systems, customer support, etc. Also, the recording of the inventory requires a large number of recording sessions and a strict recording procedure to assure the uniformity of the database (Bonafonte et al., 2006; Oliver and Szklanny, 2006; Saratxaga et al., 2006).

The high cost of the recording process limits the ability of the technology providers to produce more than a few voices for each language. A solution to this problem has been to separate the speech synthesizer engine from the inventory that defines the synthesizer's voice. Several of the public available systems allow the integration of new voices, like Festival (Black and Lenzo, 2003) and MBROLA (Dutoit et al., 1996), and some companies are willing to outsource the recording procedure in exchange for wider range of customers. Also, a properly recorded voice can be used in several systems using different technologies and have a lifespan longer than the synthesizer engines.

In this paper we describe our experience and the solutions that we have adopted to record several speech inventories in European Portuguese. These inventories were designed to be integrated in the wide range of applications produced by the companies of the *Tecnovoz* consortium.

1.1. The *Tecnovoz* Project

The Tecnovoz project is a join effort to disseminate the use of spoken language technologies in a wide range of different domains. The project consortium includes 4 research centers and 9 companies specialized in areas such as banking, health systems, fleet management, security, media, alternative and augmentative communication, computer desktop applications, etc.

To meet the goals of the project 13 demonstrators are being developed using 9 speech technology modules. Two of these modules are related with speech output: one module for domain specific speech synthesis and another for synthesis with unrestricted input. The first module will be used, for example, in banking applications were almost natural quality can be achieved by a proper match between the inventory and the desired output sentences. As an example of synthesis with unrestricted vocabulary, one of the demonstrators is a dictation machine that provides oral feedback to the user.

We have adopted a single system to handle both requirements. The domain adaptation is done at speech inventory level. The inventory can have a wide or narrow coverage of the language. By using an inventory with very large number of carefully selected samples of a restricted domain, a very high quality can be achieved for sentences in that domain. A more general purpose system can use an inventory with a wider coverage but with fewer examples for each domain.

By working together with the companies involved in the *Tecnovoz* project, we were able to create customized voices to fulfill the requirements of each particular application.

1.2. Organization of the Paper

In this paper we will start by the describing the procedures for collecting text and selecting sentences for the recording prompts. After that we will describe the methodology that we have followed for selecting the speakers. Next we will describe the details of the recording process, namely the techniques to maintain the same conditions and voice characteristics during the multiple recording sessions. Finally we describe the procedures to control the quality of the recordings, during and after the recording sessions, so that the sentences with problems could be re-recorded. We will conclude with some conclusions and future work.

2. Design of the Recording Prompts

Two distinct approaches were taken to select the text prompts to be recorded: one for specific domains and another to cover the acoustic patterns observed in the general use of the language.

2.1. Recording Prompts for Specific Domains

The domain specific prompts must not only cover the most frequent words in the domain but they must also provide a good coverage of the prosodic contexts in which they appear. To achieve this goal we started by collecting domain specific text corpora with the help of our industrial partners of the *Tecnovoz* consortium. A frequency-dependent approach was then used to decide which words should be included in the domain's word list (Ziegenhain et al., 2003). With this word list a greedy selection algorithm was used to select a representative sub-set of the sentences in the text corpora (Johnson, 1973; Chevelu et al., 2007). Due to the need for modelling common multi-word expressions that often give rise to very peculiar co-articulation phenomena, the candidate prompts were selected in order to cover the most frequent word pairs and word tri-grams. Since not all combinations exist on the text corpus, the greedy algorithm stops when a predefined coverage is achieved. Then, additional sentences are manually designed in order to cover the relevant words in appropriate contexts, if such words and contexts were not found in the automatically selected prompts.

2.2. Recording Prompts for Open Domain

The coverage of the more general use of the language cannot be achieved at the word level, as the number of words in a language is virtually infinite. Therefore, the candidate prompts must be represented by smaller sized acoustic units. We used three levels of representation: syllables, triphones and diphones. These levels make up finite sets and can carry information that spans from the phonetic level up to the prosodic level.

The creation of the text corpus for sentence selection was largely inspired by the language resource specification used in the TC-STAR project (Bonafonte et al., 2006) and took several steps. We started by taking a subset of the *WEBNEWS-PT* corpus, a text collection effort that started in 1997 (Neto et al., 1997), comprising about 1,300,000 articles from four newspapers, three generic newspapers and a sports newspapers, during the years 2003 and 2004. Those articles gave rise to around 3,200,000 sentences by using a set of Festival-embedded text analysis tools for European Portuguese. The corpus contains a total of about 70,000,000 million words and 420,000 distinct words. Many of these are foreign words, names, acronyms and other non-standard words. In order to have a proper sub-word selection scheme we need to have a very high confidence in the estimated phone sequence for every sentence. For this reason we discarded all the sentences that contained words not included in our manually corrected pronunciation lexicon. The text corpus was this way reduced to around 400,000 sentences. The sentence selection was again performed by means of a greedy algorithm aiming at covering tokens at the three selected levels: syllables, triphones and diphones. Since we cannot achieve a total coverage of them with a finite set of prompts (the number of syllables and triphones is prohibitively large for that), weighting factors are used to speed up the coverage of some levels at the expense of others. We tuned those factors in order to start by optimizing the diphone coverage. As in the specific domains case, the greedy algorithm stops after a pre-defined coverage is achieved.

2.3. Additional Recording Prompts

Although a full diphone coverage assures the possibility of synthesizing all the words in the language, the use of concatenation boundaries inside words usually has some impact in the quality of the system. A way to solve this problems is to add to the inventory some recordings covering some common lexical items.

In applications for children, for example, it is very common the use of verbs in the first and second person. These words are poorly covered in a newspaper based text corpus. Therefore, the manually designed prompts should account for several features, namely, a list of the most frequent verbs in European Portuguese in both first and second persons. Hence, we computed the frequency of occurrence of each verb lemma in a corpus of around 1,600,000 newspapers' articles, based on the results of a morpho-syntactic analysis tool (Ribeiro et al., 2003). A special set of recording prompts were then produced with those most frequent verbs.

There are many other cases of common words that are not covered in such a corpus but that can be needed in certain applications, like phone numbers, economic terms, currencies, computer science terms, some foreign names and expressions, typical dishes, touristic attractions, or even countries and their capitals. Sentences were manually created to provide coverage for these items.

Since speech synthesizers are currently being used in dialogue systems and in speech-to-speech translation systems, we also added the possibility of including in the inventory sentences transcribe from human dialogues. For this purpose we used the transcriptions of the CORAL map task corpus (Trancoso et al., 1998). A greedy algorithm was used to select a subset of representative sentences.

The newspaper text corpus has very few examples of interrogative sentences. To cover this we added manually designed prompts to account for all types of interrogative sentences and declarative sentences with the same lexical material as a yes/no question. This provides a variety of dis-

tinct intonational contours according to the sentence type.
The amount of data for each type of contents for each speaker is displayed on table 1. The presented values for the length of the recordings include the silence in the beginning and at the end of each utterance.

3. Speaker Selection

Due to the special nature of this project – the integration of speech technologies in a variety of different products – it was decided to record two voices of each gender.
The selection of the four speakers was based on the results of recording test sessions of several candidates. These were selected by personal contacts and through a voice talent recording studio. The 31 candidates, 18 female and 13 male, all native speakers from the Lisbon area, included both professional and non-professional speakers. Each candidate signed a contract were it was explained the purpose of the recordings and the use that could be made of the voice if selected for use in the system. The test consisted on recording a session of 600 sentences. The sentences were selected to have a good diphone coverage.
Using these recordings a synthesizer was built with each voice allowing us to evaluate not only the quality of the voice but also its use for this purpose. The decision was taken by listening to several phonetically rich prompts synthesized with a variable size unit selection voice using the recordings of each speaker. The decision criteria were:

- the reading naturalness;
- the duration of the recording session (number of repetitions);
- the capability of maintaining the voice quality during the recording session;
- the pleasantness of the synthesized voice;
- the voice ability to mask concatenations errors.

4. Recording Setup

The recordings were conducted in our own recording studio that includes a sound-proof room and a control station here supervision of the recording process took place. The equipment in the sound-proof room is:

- a Studio Projects T3 Dual Triode microphone;
- an anti-pop filter;
- a Brüel & Kjær Type 2230 microphone probe;
- an LCD monitor;
- a set of headphones;
- a web camera a small mirror on the wall.

The web camera and the mirror played an important role in helping the speaker maintaining a fixed distance to the microphones. The supervisor could control the speaker position in the beginning of each session by comparing the webcam image with pictures taken in previous sessions. The small mirror on the wall help the speakers in maintaining the position during the sessions: they were asked to check the position of their face in the mirror periodically. Also, in order to help the speaker using the same voice level and quality, in the beginning of each session the speakers listened to some recordings from previous sessions. There were also some fixed sentences in each block of recording prompts to allow the comparison of the different sessions.
In the control station, the signals from both microphones were digitized using a RME Fireface 800 digital mixing desk. A sampling frequency of 44.1 kHz and 24 bit quantization were used. The preset facility of the digital mixing desk were used to keep the adjustments for each speaker from one session to the other. The audio feedback and the supervisor instructions were also routed through the mixing desk to the speaker's headphones.

Figure 1: Recording booth. Notice the anti-pop filter in front of the microphone, the web camera on top of the monitor and the mirror on the left wall.

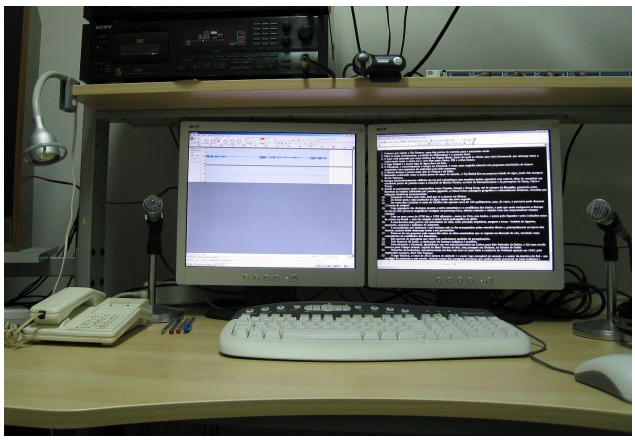

Figure 2: Control room. The display monitor on the right is a mirror of the recording booth monitor.

The control station had two display monitors, one of them being mirrored inside the sound-proof room. These monitors were used to display the recording prompts under the control of the recording supervisor. Since speaker throat relaxation and list effects have an important role in the recorded speech, recordings were done in sessions

Content Type	Utterances		Total Length		Average
	Number	%	HH:MM	%	Length
Newspaper	4200	50.8%	7:25	56.9%	6.37s
Frequent proper nouns (declarative)	432	5.2%	0:20	2.6%	2.84s
Frequent proper nouns (interrogative)	251	3.0%	0:09	1.2%	2.23s
Tourism	100	1.2%	0:14	1.8%	8.63s
Dialogue transcriptions (Coral)	600	7.3%	0:23	2.9%	2.31s
Ordinal and cardinal numbers	220	2.7%	0:09	1.2%	2.52s
Rich interrogative sentences	125	1.5%	0:03	0.5%	1.84s
Common foreign words and expressions	76	0.9%	0:10	1.4%	8.60s
Computers and Internet	35	0.4%	0:03	0.4%	5.75s
Telephone numbers	80	1.0%	0:06	0.9%	5.24s
Medical domain	1000	12.1%	2:17	17.5%	8.22s
Virtual assistants domain	288	3.5%	0:14	1.8%	2.96s
Banking domain	327	4.0%	0:40	5.2%	7.71s
Weather information domain	211	2.6%	0:14	1.8%	4.11s
Stock market domain	200	2.4%	0:26	3.3%	7.81s
Fleet management domain	115	1.4%	0:03	0.5%	1.96s
Total	8260		13:03		5.69s

Table 1: Amount of data for each type of content per speaker

of two hours with a 10 minutes interval every half hour. Each recording session produced, on average, 40 minutes of recorded speech and, except in exceptional cases, each speaker recorded only one two hour session per day. To collect a total of about 13 hours of speech per speaker we performed an average of 20 recording sessions per speaker.

5. Quality Control

The control of the recordings was done in two stages. The first was conducted during the recording session and the second by the analysis of the recordings performed after the session.

5.1. Recording Monitoring

Each recording session was monitored by two persons: a sound engineer and a recording supervisor.

5.1.1. Sound Engineer

The role of the sound engineer was to control the recording software monitoring the sound level, to start and stop the recorder and to erase the unnecessary segments. It was the responsibility of the sound engineer to detect *pops* due to excessive airflow during occlusives, usually resulting in a recording overflow, and to verify if the speaker was producing the right sound level, either by moving away from the predefined position or by starting to become tired. To speed-up the process the speaker was asked to read the sentences in sequence with short pauses between them. During this period the software recorded continuously. The pauses between sentences should have a length of, at least, 250 ms. When a problem was detected the sound engineer stopped the recorder and moved the recording cursor to the pause after the last good sentence. The recording resumed when the speaker was ready to continue.

5.1.2. Recording Supervisor

The recording supervisor role was to checked if the speaker was reading the text prompts correctly. It was necessary not only to verify if the speaker read all the words in the sentence, but if he also performed the proper pronunciation using an adequate rhythm and intonation. When the supervisor detected something wrong she asked the sound engineer to stop the recording software and told the speaker what was not right. When the speaker had doubts in the way to read a specific sentence, it was the role of the recording supervisor to clarify it. One of the most challenging sessions involved reading medical terms necessary for one of the domain-specific voices. We had to play the recordings of a medical doctor to help the speakers produce the correct pronunciation.

5.2. Analysis of the recordings

After the recording session, the first step was to segment the recordings of each session into separate audio files. Each file contains a single utterance with a margin of silence between 100 and 200 ms in the beginning and at the end. The silence included in the recordings was the silence captured by the microphone and can be used to measure the signal-to-noise ratio.

Variations in recording conditions were detected by comparing the average values of the MFCC parameters for each utterance (Richmond et al., 2007). By plotting the average values against the order of the recordings for each it can be seen if there were any change in the recording conditions. By comparing the values of different sessions one can also detect changes or errors in the setup for each speaker.

The next step was to phonetically segment the utterances using our own segmentation tool (Paulo and Oliveira, 2005). Although the segmentation tool was not fully adapted to each speaker, its results allowed us to performed some preliminary measurements on the recordings:

speech rate: by computing the number of syllables per minute we were able to detect variations in the speech rate;

erroneous pauses: the location of silences inside the utterances were compared with the location of the punctuation marks in the prompts.

pronunciation errors: major mismatches between the predicted pronunciation and the phone sequence produced by the segmentation tool were an indication that the recording required human analysis.

The recordings with errors were discarded and the sentences were re-scheduled to future recording sessions.

6. Conclusions and Future Work

In this paper we have detailed the main problems in the design and recording of speech databases to be used by corpus-based speech synthesizers. Among those are the collection and selection of texts related with the application domains, the selection of appropriate speakers, all the necessary techniques for assuring and maintaining the same quality during the multiple recording sessions and the criteria for the acceptance of the resulting recordings. The *Tecnovoz* project adopted several strategies to address these problems that resulted in a speech database with 4 speakers with a total of 6 hours of speech per speaker. These recordings are group into application domains that can be combined to generate inventories for different speech synthesis applications.

Although the automatic segmentation of the recordings has helped in locating several problems, this task demands further improvements in this tool. In one hand we need to increase the performance of the baseline segmentation tool with limited adaptation to the speaker, since the adaptation data is not available in the beginning of the recording process. On the other hand the segmentation process cannot always rely on the predicted phonetic sequence to perform the alignment. Although our tool already allows some alternative pronunciations, it should allow more variability without compromising its accuracy.

Another difficulty that we are working on is the detection of prosodic ruptures when there is no pause in the signal. The correct location of the phrasing boundaries is an important factor in the assessment of the rhythm and intonation produced by the speaker.

7. Acknowledgements

This work was funded by PRIME National Project TECNOVOZ number 03/165. INESC-ID is funded by Fundação para a Ciência e Tecnologia (FCT).

8. References

Alan W. Black and Kevin A. Lenzo, 2003. *Building Synthetic Voices*.

A. Bonafonte, H. Höge, I. Kiss, A. Moreno, U. Ziegenhain, H. van den Heuvel, H.-U.Hain, X.S.Wang, and M. N. Garcia. 2006. Tc-star: specifications of language resources and evaluation for speech synthesis. In *LREC-2006: Fifth International Conference on Language Resources and Evaluation*, pages 311–314, Genoa, Italy, May.

Jonathan Chevelu, Nelly Barbot, Olivier Boeffard, and Arnaud Delhay. 2007. Lagrangian relaxation for optimal corpus design. In *Proceedings of the 6th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*, pages 211–216. ISCA.

T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. V. der Vrecken. 1996. The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Proc. ICSLP '96*, volume 3, pages 1393–1396, Philadelphia, PA.

David S. Johnson. 1973. Approximation algorithms for combinatorial problems. In *STOC '73: Proceedings of the fifth annual ACM symposium on Theory of computing*, pages 38–49, New York, NY, USA. ACM.

Joao P. Neto, Ciro A. Martins, Hugo Meinedo, and Luis B. Almeida. 1997. The design of a large vocabulary speech corpus for the portuguese. In *Proc. Eurospeech '97*, pages 1707–1710, Rhodes, Greece.

Dominika Oliver and Krzysztof Szklanny. 2006. Creation and analysis of a polish speech database for use in unit selection synthesis. In *LREC-2006: Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy, May.

Sérgio Paulo and Luís C. Oliveira. 2005. Generation of word alternative pronunciations using weighted finite state transducers. In *Interspeech'2005*, pages 1157–1160. ISCA, September.

Ricardo Daniel Ribeiro, Luís C. Oliveira, and Isabel Trancoso. 2003. Using morphossyntactic information in tts systems: Comparing strategies for european portuguese. In *PROPOR'2003 - 6th Workshop on Computational Processing of the Portuguese Language*, Lecture Notes in Artificial Inteligence, pages 143–150. Springer-Verlag, Heidelberg, June.

Korin Richmond, Volker Strom, Robert A J Clark, Junichi Yamagishi, and Sue Fitt. 2007. Festival multisyn voices for the 2007 blizzard challenge. In *Proc. Blizzard Challenge Workshop (in Proc. SSW6)*. ISCA.

I. Saratxaga, E. Navas E., I. Hernaez, and I. Luengo. 2006. Designing and recording an emotional speech database for corpus based synthesis in basque. In *LREC-2006: Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy, May.

Isabel Trancoso, M. Céu Viana, Inês Duarte, and Gabriela Matos. 1998. Corpus de diálogo coral. In *PROPOR'98 - III Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada*, November.

U. Ziegenhain, H. Hoge, V. Arranz, M. Bisani, A. Bonafonte, N. Castell, D. Conejero, E. Hartikainen, G. Maltese, K. Oflazer, A. Rabie, D. Razumikin, S. Shammass, and C Zong. 2003. Specification of corpora and word lists in 12 languages. Technical Report 1.3, Siemens AG, April.