Database for multi-pitch tracking

Student Project

by

Gregor Pirker

Signal Processing and Speech Communication Laboratory
at Graz University of Technology

Head: Kubin, Gernot, Univ.-Prof. Dipl.-Ing. Dr.techn.

Assessor: Pernkopf, Franz

Supervisor: Pernkopf, Franz

Graz, May 2010

Abstract

This report introduces a speech database for multi-pitch tracking and describes the main steps of its development. The database contains the audio recordings and laryngograph signals of 20 English native speakers as well as the extracted pitch signals as a reference. The text material consists of 2342 “phonetically rich” sentences which are taken from the existing TIMIT corpus [2]. Each of them was at least once read by a female and a male speaker. In total this database consists of 4720 recorded sentences. All recordings were supervised and carried out at the recording studio of the Institute of Broadband Communication at Graz University of Technology. The speech recording program and the appropriate recording setup will be described in detail in this report. Furthermore a snapshot of the data is provided by depicting some examples to show the connections between speech signals and pitch tracks. As this project will be continued, this report gives also a preview of the next steps like documentation, validation and distribution.

Contents

1 Introduction

A pitch tracking algorithm tries to estimate the pitch or the fundamental frequency of speech which is important for applications like e.g. one-channel blind source separation. The Signal Processing and Speech Communication (SPSC) Laboratory at Graz University of Technology developed such a pitch tracking algorithm for multiple speakers talking simultaneously.

The aim of this project is to provide a database as a basis for evaluation and comparison of various multi-pitch tracking algorithms.

In terms of cooperation with international research in the area of speech analysis and pitch tracking this database consists of spoken English utterances, read by native speakers. Another major aspect in this context is to provide the data with thorough documentation.

Chapters 2 and 3 of this report deal with the preparation and the development of the database respectively, which were carried out by means of the "ready-to-use solutions" in [1]. Therefore, we use most of the technical terms from this book and follow the specified order. All special properties and utilities concerning this speech corpus are described in detail. Chapter 4 gives some examples of the recorded data and particularly explains extracting the pitch contour from the laryngograph signals.

The final validation, documentation and publication of the database on a website will be part of the subsequent diploma thesis.

2 Corpus Specifications

The first step of the production of a speech corpus is to specify all the desired features. At the beginning this chapter provides a list of definitions of important terms concerning speech corpora taken from [1]. The main part contains the description of the general and the technical specifications (as suggested in [1]) of our database. In addition one important feature of this database is discussed in detail: The spoken sentences are taken from the existing TIMIT corpus [2].

2.1 Terms and Definitions

Below some important definitions of technical terms (taken from [1]) in the context of this project are provided:

> "Speech Corpus = physical time signals, in most cases sound pressure or other measurable time signals recorded from the act of speaking, together with an associated set of annotations, meta data and documentation stored on a digital medium.
> Validation = the (formal) check of a speech corpus with regard to its pre-defined specifications.
> Evaluation = a qualitative assessment of a corpus with regard to its usability in a certain task or development scenario.
> Specification = the fixed technical description of a speech corpus with regards to all of its features (including annotations, meta data and documentation).
> (File) Format = standardized or specified format of digital signal and symbolic (annotations, meta data) data.
> Annotation = discrete (categorized) description associated with a physical signal (coding). Usually consists of a closed set of symbols and a scheme to link these symbols to either points in time or segments in time.
> […]
> Prompt = speech item (word, phrase or sentence) presented to a speaker. A prompt list or prompt corpus is a collection of prompts that define the spoken content of the corpus.
> Spoken Content = what was spoken in a speech corpus.
> Meta data = data about data. In this report the term meta data is restricted to three types: recording protocols, comments and speaker profiles." [1]

2.2 General Specifications

This first part of the specifications defines the main properties of this speech corpus as suggested in [1].

Speaker Profiles

The speakers for our recordings are English native speakers or bilingual speakers with a gender distribution of 50:50 (%).

Number of Speakers

At least 20 speakers were required.

Spoken Contents

For the contents of the corpus the sentences from the existing *TIMIT* database were used. This database will be introduced in chapter 2.4.

Speaking Style

The speaking style was read speech.

Recording Setup

For this database a supervised on-site recording was carried out. Chapter 3.3 gives more detailed information about that.

Acoustical environment

As recording room the recording studio at the Institute of Broadband Communication at Graz University of Technology was chosen. Each speaker had to be recorded on-site.

Background Noise

The acoustical background consisted only of the hum of the recording PC which was located 2 m from the head of the speaker, separated by an absorbing wall. Chapter 3.3.1 describes the acoustical environment in detail.

Script

The speakers read sentences from a screen while not changing position.

Microphone

To record the speech signal, the Headset *AKG HC 577 L* was used.

Laryngograph

In addition the speakers had to wear electrodes of the *Portable Laryngograph®*.

2.3 Technical Specifications

According to [1] the technical specifications give an overview of the possible categories and values of all signals and symbolic data.

Sampling rates

Both the microphone signal and the laryngograph signal are sampled at 48 kHz.

Sample Type and Width, Encoding

The type of encoding is signed PCM, 16 bits and the byte order type is little endian.

Number of Channels

Two channels were recorded. The left channel was used for the microphone, the right channel for the laryngograph.

Signal File Formats

The WAV format (stereo; L: microphone, R: laryngograph) serves as the file format for the recorded signals.

Meta Data File Formats

The database provides two TXT files with meta data, the recording protocol and the speaker profile.

Meta Data Contents
The recording protocol includes:

- Session ID
- Speaker ID
- Recording Date
- Environmental conditions
- Technical recording conditions
- Microphone
- Laryngograph
- Recording device
- Technical specifications of recorded signals
- Placement and distance to microphone
- Name or ID of the recording

Type of Prompting
Supervisor presence
Type of speech
Free comments (if necessary)

The speaker profile contains:

Speaker ID
Age
Sex
Mother tongue/Home country

The description of the corpus structure, the terminology, the distribution Media, and the documentation will be part of the subsequent diploma thesis.

2.4 Spoken Sentences – The TIMIT Corpus

For the spoken content we used the sentences from the *TIMIT* corpus [2]. This corpus "has been designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. […] Text corpus design was a joint effort among the Massachusetts Institute of Technology (MIT), Stanford Research Institute (SRI), and Texas Instruments (TI)." [2] In the text material of the TIMIT database three different types of "phonetically rich" sentences can be found which are described below.

2.4.1 Text material

The spoken content in the *TIMIT* prompts consist of two "dialect sentences" (named: sa + sentence-number) "to expose the dialectal variants of the speakers", 450 "phonetically-compact" sentences (sx + sentence-number) "to provide a good coverage of pairs of phones, with extra occurrences of phonetic contexts thought to be either difficult or of particular interest", and 1890 "phonetically-diverse" sentences (si + sentence-number) "to add diversity in sentence types and phonetic contexts". [2]

Here are some examples:

Dialect sentence (sa1)

```
She had your dark suit in greasy wash water all year.
```

Phonetically-compact sentence (sx409)

```
Eating spinach nightly increases strength miraculously.
```

Phonetically-diverse sentence (si1291)

```
They should live in modest circumstances, avoiding all conspicuous
consumption.
```

Table 1 illustrates the distribution of sentences among speakers with respect to this project: The two "dialect sentences" were read by all 20 speakers. Additionally each speaker read 45 of the "phonetically-compact" sentences and 189 of the "phonetically-diverse" sentences. Hence each set of these sentences was spoken by two speakers, one female and one male speaker.

Sentence Type	#Sentences	#Speakers	Total	#Sentences/ Speaker
Dialect (sa)	2	20	40	2
Compact (sx)	450	2	900	45
Diverse (si)	1890	2	3780	189
Total	**2342**		**4720**	**236**

Table 1: Distribution of the three types of phonetically rich sentences

3 Recording Setup

Contrary to the last chapter were we determined all required specifications for our database in a formal manner, this section describes the practical implementation.

During the first preparation phase the main issue was to implement a setup as well as recruiting speakers including legal aspects. The recruitment and recording had to be carried out in parallel.

This part of the report gives a quick insight in the preparation and in the recording routine and provides a detailed description of the recording setup. In this context recording setup means both, acoustical environment and technical setup. Furthermore, one of our main tools concerning the technical setup that will be introduced is a recording software called SpeechRecorder, developed by the Institute of Phonetics and Speech Processing of Ludwig-Maximilian University in Munich, Germany. [3]

3.1 Preparation of Recording

After defining the corpus specifications, preparation was needed before the recording session could be started. Concerning the speech corpus recording the most important issues are listed in the following:

Legal Aspects

Each speaker had to be advised about the purpose of the recording, data protection and anonymity and had to sign a declaration of allowance to use the recordings and some (insensible) data. Furthermore this form included a confirmation of receiving financial compensation.

The collected data were age, sex and home country.

(Plan) Recruitment

Although the SPSC offered money (20 €) this part of the preparation was difficult and time-consuming. However, after posting in appropriate newsgroups and advertising at proper institutions and associations the first candidates responded. The best asset in the recruitment were the recruited speakers themselves.

Closely related to the procedure of the recruitment was the (time-) scheduling of the recording sessions and the studio reservation.

Pre-test

A pre-test is necessary to eliminate all bugs in the procedure. Beside some smaller trials which did not take place in the studio we made the last test on-site and exactly like planed in the recording session.

3.2 Recording procedure

To obtain feasible recordings in every session, it is a good advise to plan the recording routine thoroughly. The following checklist is the procedure which we used in the course of the project:

1) After meeting speakers the legal aspect and the data collection (see chapter 3.1) were discussed and some instructions and explanations were given.

2) Finding the right position of the laryngograph-electrodes and fixing the headset sometimes can be difficult. Beards, hairdos, speakers moving too much or touching and displacing the electrodes are some reasons which made supervision during the recording necessary. Before the recording the signals need to be tested again.

3) The main recording phase took about one hour. It was carried out sentence by sentence so repetitions could be done easily. Since this was a very discomforting situation to most of the speakers, we had to include breaks, which means that step two (replacing electrodes and headset and testing signals) had to be performed again.

4) At the end the speakers received money and were encouraged helping recruiting people.

3.3 Recording setup

> "Basically the recording setup defines the acoustical characteristics of the resulting corpus and therefore the usability of the data for certain applications or investigations" [1].

The appropriate setup for this speech corpus production was a supervised on-site recording in an appropriate well defined recording studio.

3.3.1 Acoustical Environment

As described in the corpus specifications the recording was done in the recording studio at the Institute of Broadband Communication at Graz University of Technology. Figure 3.1 shows our setup with the two chairs and screens; on the right for the supervisor and on the left for the speaker. The supervisor controlled and monitored the recording procedure (chapter 3.2) with

the help of the recording software (chapter 3.3.3) and the headphones. The speaker was equipped with the headset and the neck band with the laryngograph – electrodes. The speaker reads the displayed sentences from the screen.

Fig. 3.1: Recording Setup

Since the recording Laptop produced background noise, it was separated from the speaker by an absorbing wall.

3.3.2 Technical Setup

In Figure 3.1 also the technical setup for this task is shown. It consisted of a laptop equipped with a particular speech recording program (chapter 3.3.3), a firewire recording interface, headphones, the laryngograph including the electrode-neckbands and the headset.

The technical specifications in chapter 2.3 include details about sampling rate, resolution and also list the main devices necessary to be able to reproduce the corpus. In the following, the complete list of devices is given:

Recording PC: IBM Laptop, Type 2366
Recording interface: Presonus Firebox

Microphone: AKG HC 577 L, Headset
Laryngograph: *Portable Laryngograph®*

3.3.3 Recording Software

The special corpus recording software *SpeechRecorder* [3] was developed by the Institute of Phonetics and Speech Processing of Ludwig-Maximilian University in Munich, Germany.

> "*SpeechRecorder* is an application for script-driven speech, audio, and signal recordings. Its main features are:
> - platform independence
> - automatic and manual recording progress
> - local and remote recordings via the Internet
> - number of recording channels dependent only on the audio hardware
> - speaker and supervisor views on multiple screens
> - full Unicode text, image and audio prompts
>
> […]
> *SpeechRecorder* organizes recordings in projects. A project is a combination of a speaker database, a set of recording scripts, and a set of recording sessions. A recording session consists of an individual speaker, a recording script, the selected recording settings, and a directory into which the recorded files are written." [3]

The project of *SpeechRecorder* contains a speaker database, a directory for the audio recordings, a project configuration file, a sample recording script, and the recording script DTD. [3]

The following passages describe the *SpeechRecorder* with our settings for recording the speech and the larynx signals. Basically, there are many more possibilities to adapt this software for any other speech recording application.

Multiple Displays

Since the recordings had to be supervised we used two screens, one for the speaker (Fig. 3.2), one for the supervisor (Fig. 3.3). The speaker view shows the current sentence and a traffic light to indicate the recording phase.

> "Each recording is performed as a sequence of phases.
> […]
> IDLE no recording, red light, prompt item is only displayed if the attribute promptphaseis set to idle.
> PRERECORDING recording, yellow light, modal prompt item display.
> RECORDING recording, green light, active prompt item display.
> POSTRECORDING recording, yellow light, modal prompt item display." [3]

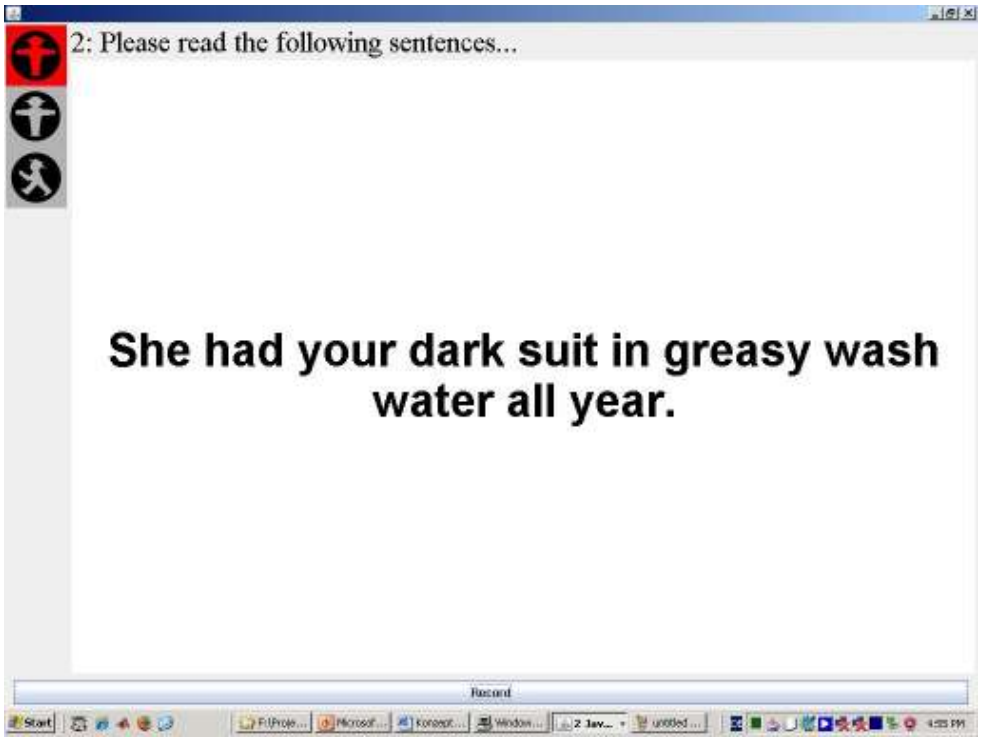

Fig. 3.2: Speaker view

The supervisor view additionally includes a level meter (bottom left), the signal display and the list of sentences.

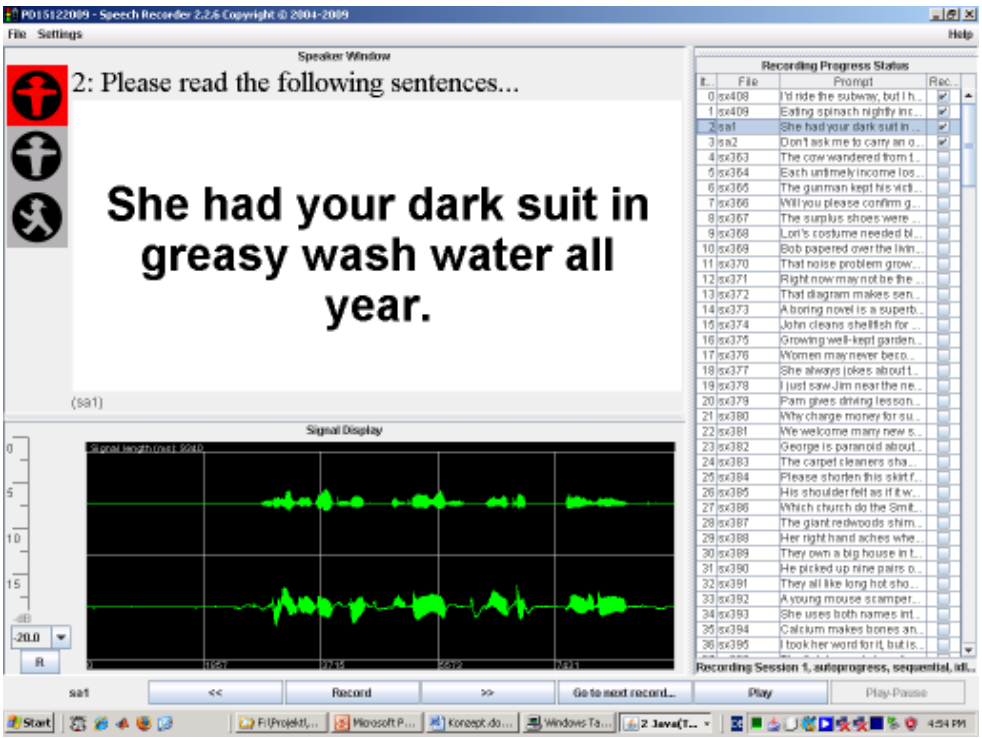

Fig. 3.3: Supervisor view

Script

> "A script specifies which items are to be recorded. A script consists of two parts, a header containing meta-data items, and the recording script proper. The recording script is divided into sections. A section is an organizational unit that specifies the presentation order, and progress mode for the recording items it contains. A recording item consists of the instructions, the prompt item, and a comment. Instructions and comments are optional." [3]

For this project, section one of the following script (`"Introduction and Testing"`) included two random sentences which were used to check the setup, adjust the signal level and familiarize the speaker with the system (recording procedure, reading-emphasis, reading-speed, …). The prompt items of section two (`"Recording Session 1"`) provided the phonetically rich sentences.

The following script shows the header and the two sections with one example-sentence each.

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
<!DOCTYPE script SYSTEM "SpeechRecPrompts.dtd">

<script id="1m1">
      <metadata>
            <key>
                  DatabaseName
            </key>
            <value>
                  SPSC - Pitchtracking Database
            </value>
            <key>
                  ScriptName
            </key>
            <value>
                  TIMIT Prompts - Part 1
            </value>
            <key>
                  ScriptAuthor
            </key>
            <value>
                  Gregor Pirker
            </value>
            <key>
                  EmailAuthor
            </key>
            <value>
                  No Email
            </value>
      </metadata>

      <recordingscript>


            <section name="Introduction and Testing" order="sequential"
speakerdisplay="yes" mode="autoprogress" promptphase="idle">

            <recording  prerecdelay="2000" recduration="60000"
postrecdelay="500" itemcode="sx409">
                  <recprompt>
                        <mediaitem mimetype="text/UTF-8">
                              Eating spinach nightly increases strength
miraculously.
                        </mediaitem>
```

```
                  </recprompt>
                  <reccomment>
                        (sx409)
                  </reccomment>
            </recording>

            </section>



            <section name="Recording Session 1" order="sequential"
speakerdisplay="yes" mode="autoprogress" promptphase="idle">

            <recording  prerecdelay="2000" recduration="60000"
postrecdelay="500" itemcode="sa1">
                  <recinstructions mimetype="text/ISO-8859-1">
                        Please read the following sentences...
                  </recinstructions>
                  <recprompt>
                        <mediaitem mimetype="text/UTF-8">
                              She had your dark suit in greasy wash water
all year.
                        </mediaitem>
                  </recprompt>
                  <reccomment>
                        (sa1)
                  </reccomment>
            </recording>



            <recording  prerecdelay="2000" recduration="60000"
postrecdelay="500" itemcode="si453">
                  <recprompt>
                        <mediaitem mimetype="text/UTF-8">
                              Everything went real smooth, the sheriff said.
                        </mediaitem>
                  </recprompt>
                  <reccomment>
                        (si453)
                  </reccomment>
            </recording>



            </section>
      </recordingscript>
</script>
```

4 Database

The content of the SPSC database consists of 4720 audio recordings from 20 different English native speakers (composition see chapter 2.4.1, Table 1) and just as much corresponding laryngograph signals. Additionally it includes the pitch contours extracted out of these laryngograph signals to provide reference pitch values. In this chapter one example-sentence is used for the purpose of illustration.

Example-sentence (sx122):

`Encyclopedias seldom present anecdotal evidence.`

4.1 Recorded Signals

The signals of this corpus recorded by the microphone and the laryngograph were digitized at 48 kHz and 16 bit resolution. Figure 4.1 depicts the characteristic and the coherence between corresponding waveforms of the word ‘encyclopedias’ from sentence (sx122). Note that these are the raw signals without any post-processing.

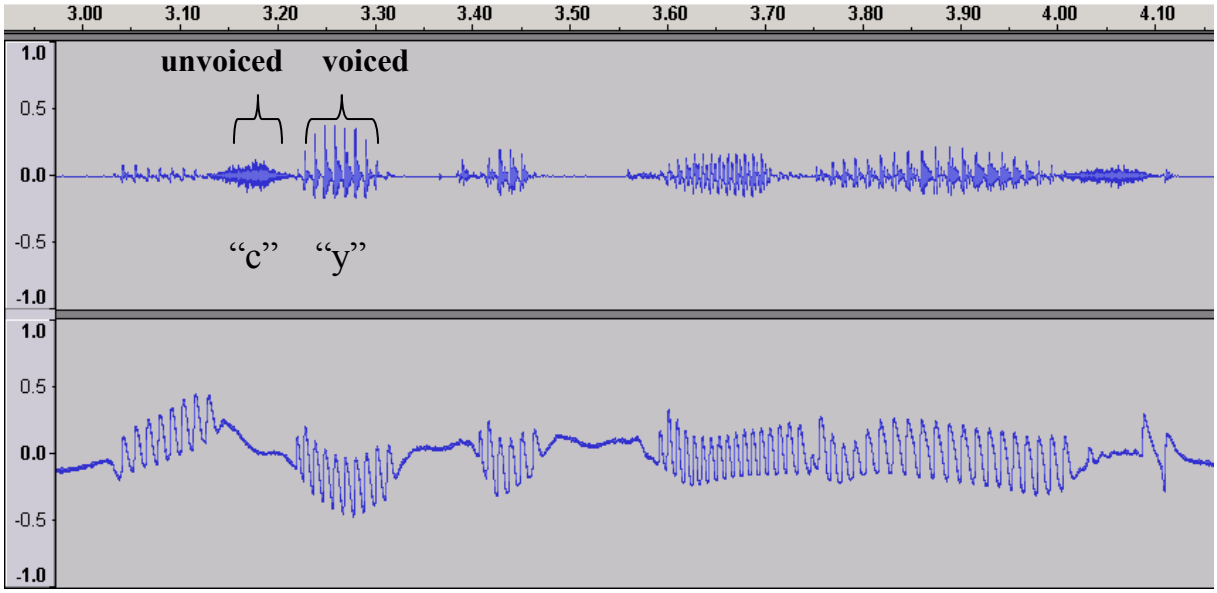

Fig. 4.1: The top figure shows the recorded microphone waveform over time, whereas the bottom waveform represents the laryngograph signal of the word ‘encyclopedias’. Speech includes voiced and unvoiced parts.

Microphone Signals

In general speech signals are non stationary and time varying. They can be voiced (vowels), unvoiced (fricatives, plosives, …), or a mixture of these two characteristics (voiced consonants) [4]. In the microphone signal in Fig. 4.1 (top) we marked the unvoiced "c" and the voiced "y".

Laryngograph Signals

The laryngograph signal is produced as follows:

> "A small high-frequency electric current is passed through the larynx by a pair of electrodes that are pressed against the neck at the position of the larynx from both sides. The opening and closing of the glottis during each pitch period cause the laryngeal conductance to vary; thus the high-frequency current is amplitude modulated. In the receiver the current is demodulated and amplified. Finally, the resulting signal is high-pass filtered to remove unwanted low-frequency components due to vertical movement of the larynx." [5]

The laryngogram in Fig. 4.1 (bottom) shows the raw signal including the high frequency part due to the vocal folds vibrations and the low frequency part mainly caused by (vertical) larynx movement. Before pitch extraction is performed the low frequency part is removed by a highpass filter.

4.2 Reference Signals

Pitch extraction in this context means fundamental frequency estimation of the laryngograph waveform. This waveform has a quasi-periodic shape which can be related to the vibration of the vocal folds. As mentioned in the previous chapter the recorded laryngograph signals need to be highpass filtered. Finally the RAPT – algorithm [6] was used by means of the speech analysis tool wavesurfer (by Kåre Sjölander and Jonas Beskow) and for practical reasons also by a pitch extraction (Tcl-) script to extract the pitch from filtered laryngograph signals. Figures 4.2 shows both the unfiltered and the highpass filtered laryngograph signal of sentence (sx122) spoken by a male voice. Since the raw signal in this case was already smooth and band limited, filtering did (almost) not affect the signal. Therefore the green curve (filtered) covers the red curve (unfiltered). In contrast Figure 4.3 shows the same example but this time sentence (sx122) was spoken by a female voice. The pitch contour achieved from the unfiltered signal contains a lot of outliers. The filter used in this example only partly improved the pitch contour by eliminating some of the outliers (orange circle), additionally the pitch was also distorted (yellow cycle).

For the two examples filtering was achieved by means of the speech analysis program Praat (by Paul Boersma and David Weenink, University of Amsterdam) with the following filter settings:

Male speech: Cut-off frequency fc = 25 Hz, Smoothing: 50 Hz
Female speech: Cut-off frequency fc = 100 Hz, Smoothing: 70 Hz

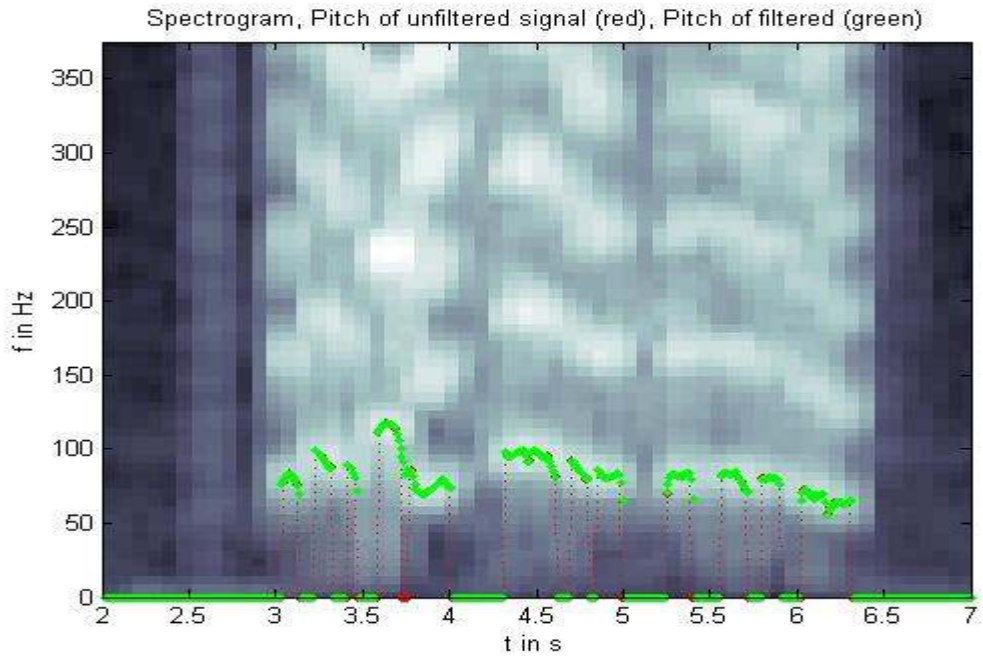

Fig. 4.2: Pitch tracks of a male speaker reading sentence (sx122); green: extracted from the high pass filtered laryngograph signal, red: extracted from the unfiltered laryngograph signal. In this case filtering was unnecessary.

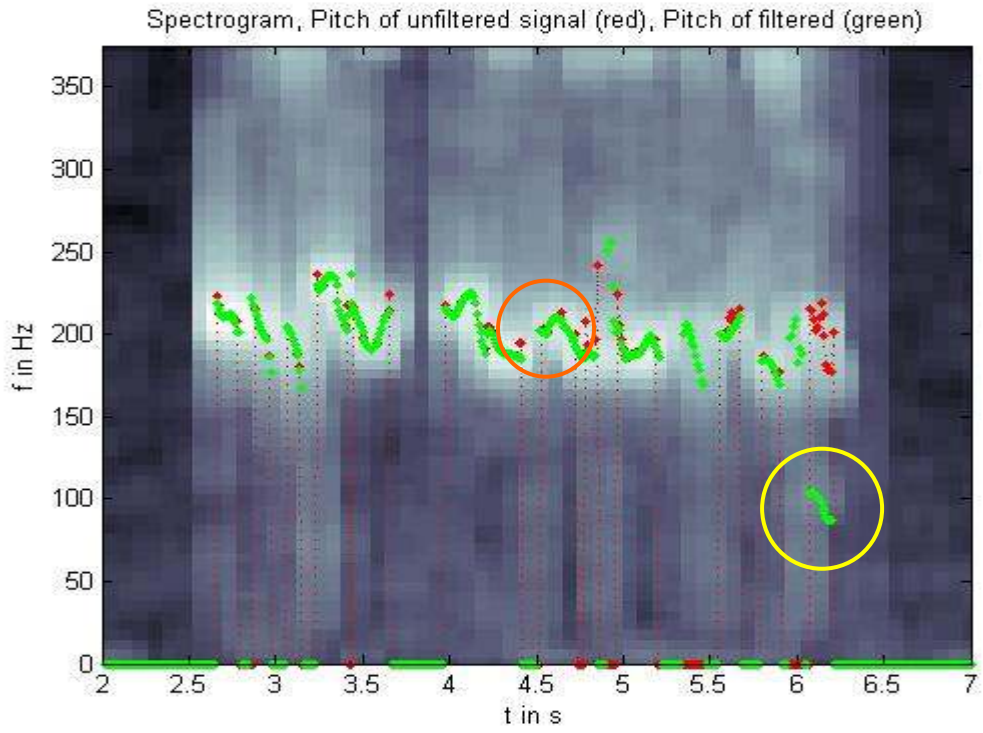

Fig. 4.3: Pitch tracks of a female speaker reading sentence (sx122); green: extracted from the high pass filtered laryngograph signal, red: extracted from the unfiltered laryngograph signal. This is an example of an insufficient filter: Outliers are only partly eliminated (orange circle) additionally the pitch track was also distorted (yellow circle).

Multi - Pitch Signals

For the evaluation of multi-pitch tracking algorithms two or more recordings have to be mixed at 0 dB. The resulting pitch track can then be compared with the two corresponding reference signals. More details about multi-pitch tracking are given in [7].

5 Conclusion

The recording of the SPSC multi-pitch tracking database is completed. It consists of 4720 audio recordings from 20 English native speakers and just as much laryngograph signals and reference pitch tracks. The spoken content are the 2342 “phonetically rich” sentences from the existing TIMIT corpus [2] each of them was read by both female and male speakers. All recordings were supervised and carried out on-site at the recording studio of the Institute of Broadband Communication at Graz University of Technology. This document also reported the steps of production as well as the subsequent signal processing and pitch extraction by means of some examples. Currently, the focus is on improved signal processing methods for accurate reference pitch extraction. Future work consists of: documentation, final validation and publishing the database on a website. This will be done in the subsequent diploma thesis. Additionally, we aim at evaluating the SPSC multi-pitch tracking algorithm [7], using this database and compare the result to the algorithm of Wu et al. [8]. The SPSC database will enable to evaluate and compare algorithms for multi- pitch tracking.

6 References

[1] Schiel F., Draxler Ch.: *Production and Validation of Speech Corpora*. Bastard Verlag, München, 2003

[2] Garofolo J. S., Lamel L. F., Fisher W. M., Fiscus J. G., Pallett D. S., Dahlgren N. L., Zue V.: *readme.doc, update 10/12/90. TIMIT* Online documendation

[3] Draxler Ch.: *SpeechRecorder Quick Start and User Manual.* Institut für Phonetik und Sprachverarbeitung, Universität München

[4] Vary P., Heute U., Hess W.: *Digitale Sprachsignalverarbeitung*. B.G. Teubner, Stuttgart, 1998

[5] Benesty J., Sondhi M. M., Huang y.: *Springer handbook of speech processing*. Springer Verlag, Berlin, 2007

[6] Talkin D.: *A robust algorithm for pitch tracking*. Elsevier Science B.V., 1995

[7] Wohlmayr M., Pernkopf F.: *Finite Mixture Spectrogram Modeling for Multipitch Tracking Using A Factorial Hidden Markuv Model*. Interspeech, 2009

[8] M., Wang D. and Brown G.J.: *A multipitch tracking algorithm for noisy speech*. IEEE Transactions On Speech and Audio Processing, 11(3):229-241, 2003