Graz University of Technology

The Pitch-Tracking Database from Graz University of Technology

Author: Gregor Pirker, Michael Wohlmayr, Stefan Petrik, Franz Pernkopf
Date: Graz, August 22, 2012
Rev.: alpha 1.1

Abstract

The Pitch Tracking Database from Graz University of Technology (PTDB-TUG) is a speech database for pitch tracking that provides microphone and laryngograph signals of 20 English native speakers as well as reference pitch trajectories. Each subject read 236 out of 2342 phonetically rich sentences from the existing TIMIT corpus [2]. The text material was selected such that each sentence was spoken by at least one female and one male speaker. In total this database consists of 4720 recorded utterances. All recordings were carried out on-site at the recording studio of the Institute of Broadband Communications at Graz University of Technology. In this report an exposition of all the properties and a description of the main steps of production are presented.

Contents

1 Introduction

A pitch tracking algorithm usually estimates the pitch or the fundamental frequency of human speech or music signals. The Signal Processing and Speech Communication (SPSC) Laboratory at Graz University of Technology developed such an algorithm for multiple speakers talking simultaneously [4] [5]. In the course of advancement this multi pitch tracker had to be evaluated and compared to similar algorithms by means of proper speech data. In order to achieve both, sufficient speaker dependent and speaker independent modelling, this data had to meet the following requirements:

- a substantial amount of speech data composed of phonetically rich sentences that allows for meaningful training of speaker-dependent models and
- a variety of female and male speakers such that a multi-pitch tracker can be evaluated seriously.

Since no existing database fulfilled our requirements, it was a consequent step to produce the PTDB-TUG (Pitch Tracking Database from Graz University of Technology). This database is provided on the website of the SPSC Laboratory at Graz University of Technology for research purposes in the area of speech analysis and pitch tracking. Evaluation results of the multi pitch tracker can be found in [7]. The PTDB-TUG includes signals recorded from microphone, which are supposed to be the data for algorithm testing. One can use either the provided reference signals or extract own ground truth data from the laryngograph signals if desired. At the beginning of this report the specifications of the final speech corpus are introduced. After that a closer look is taken to the profiles of the participants as well as to the spoken content, which was taken from [2]. The further sections deal with the collection of the microphone and laryngograph data and with the required post processing steps, which primarily means the extraction of the reference pitch signals. In the end an overview of the corpus structure and its terminology is provided. The whole database production process was carried out following mainly the suggestions from [1].

2 Specifications

20 English native speakers, of which 10 were female and 10 were male speakers, contributed to the PTDB-TUG. The text material consists of 2342 phonetically rich sentences, which are taken from the existing TIMIT corpus [2] and were 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 Communications at Graz University of Technology. The acoustical background consisted only of the fan noise of the recording notebook which was located 2m from the head of the speaker and separated by an absorbing wall. The speakers had to read the sentences from a screen, while being recorded by means of a headset microphone and a laryngograph simultaneously. Both, microphone signals and laryngograph signals were recorded at 48 kHz sampling rate, 16 bits resolution, with the type of encoding signed PCM and the byte order type little endian. Two channels were recorded in one stereo WAV file. The left channel was used for the microphone, the right channel for the laryngograph. The final database provides both microphone signals and laryngograph signals as single-channel WAV files. For the reference pitch data the output file of the RAPT pitch tracking algorithm [6] - an ASCII format file with the extension '.f0' - is used. This file contains a four column matrix which includes the pitch, a voicing decision, the root mean square values and the peak-normalized autocorrelation values respectively. In addition, the database provides some text files with meta data like the recording protocol, the speaker profiles and a list of the TIMIT prompts.

3 Speaker Profiles

For cooperation and comparison with international research, in particular in the area of pitch tracking, the main requirement for the participating persons was to be an English native speaker. The subjects were recruited by means of postings to newsgroups and advertisements at appropriate institutions and associations as well as word-of-mouth recommendations. Each speaker was informed about the purpose of the recording, data protection and anonymity and had to sign a declaration of allowance, in order to enable us to use the recordings and some insensible data in the database. 20 English native speakers from five different countries contributed to the database. The gender distribution is 50:50, the age varies from 22 to 48 years. Table 3.1 provides the complete speaker profiles.

Speaker ID	**Age**	**Sex**	**Home Country**	**Sentences**			**Comment**
F01	40	Female	Ireland	sa1,2	sx3-47	si453-641	
F02	25	Female	USA	sa1,2	sx48-92	si642-830	
F03	22	Female	Canada	sa1,2	sx93-137	si831-1019	
F04	26	Female	Canada	sa1,2	sx138-182	si1020-1208	
F05	48	Female	USA	sa1,2	sx183-227	si1209-1397	
F06	28	Female	USA	sa1,2	sx228-272	si1398-1586	
F07	24	Female	USA	sa1,2	sx273-317	si1587-1775	
F08	22	Female	England	sa1,2	sx318-362	si1776-1964	Error: si1912,1913
F09	22	Female	USA	sa1,2	sx363-407	si1965-2153	
F10	35	Female	USA	sa1,2	sx408-452	si2154-2342	
M01	24	Male	South Africa	sa1,2	sx3-47	si453-641	
M02	40	Male	England	sa1,2	sx48-92	si642-830	
M03	35	Male	England	sa1,2	sx93-137	si831-1019	
M04	26	Male	USA	sa1,2	sx138-182	si1020-1208	
M05	25	Male	England	sa1,2	sx183-227	si1209-1397	
M06	23	Male	USA	sa1,2	sx228-272	si1398-1586	
M07	24	Male	USA	sa1,2	sx273-317	si1587-1775	
M08	24	Male	England	sa1,2	sx318-362	si1776-1964	
M09	24	Male	Canada	sa1,2	sx363-407	si1965-2153	
M10	33	Male	USA	sa1,2	sx408-452	si2154-2342	

Table 3.1: Speaker Profiles

Note: In the course of data collection an error occurred. Hence the indicated sentences are not available.

4 Spoken Content

As text material the sentences from the existing TIMIT corpus, which is intended to be used in speech research purposes, were taken. In these prompts three different types of phonetically rich sentences can be found: There are two dialect sentences to expose the dialectal variants of the speakers, 450 phonetically-compact sentences 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 to add diversity in sentence types and phonetic contexts. A detailed description can be found in [2]. Table 4.1 shows the sentence labeling of the TIMIT prompts.

Sentence Type	**Labeling**
dialect sentences	sa`[sentence number out of {1,2}]`
phonetically-compact s.	sx`[sentence number out of {3,4,5,... ,451,452}]`
phonetically-diverse	si`[sentence number out of {453,454,455,... ,2341,2342}]`

Table 4.1: The sentence labeling of the TIMIT prompts

Sentence 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 4.2 illustrates the distribution of sentences among speakers in the PTDB-TUG: 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 of these sentences was spoken by two different speakers, once by a female and another time by a male speaker. The mapping of the speaker ID to the corresponding part of sentences is included in Table 1 and can also be found in SPEAKER-PROFILES.TXT in the documentation directory.

Sentence Type	**#Sentences**	**#Speakers**	**Total**	**#Sentences/Speaker**
Dialect sentences (sa)	2	20	40	2
Phonetically-compact s. (sa)	450	2	900	45
Phonetically-diverse s. (si)	1890	2	3780	189
Total	**2342**		**4720**	**236**

Table 4.2: Distribution of the three types of phonetically rich sentences

5 Corpus Structure and Terminology

5.1 Structure

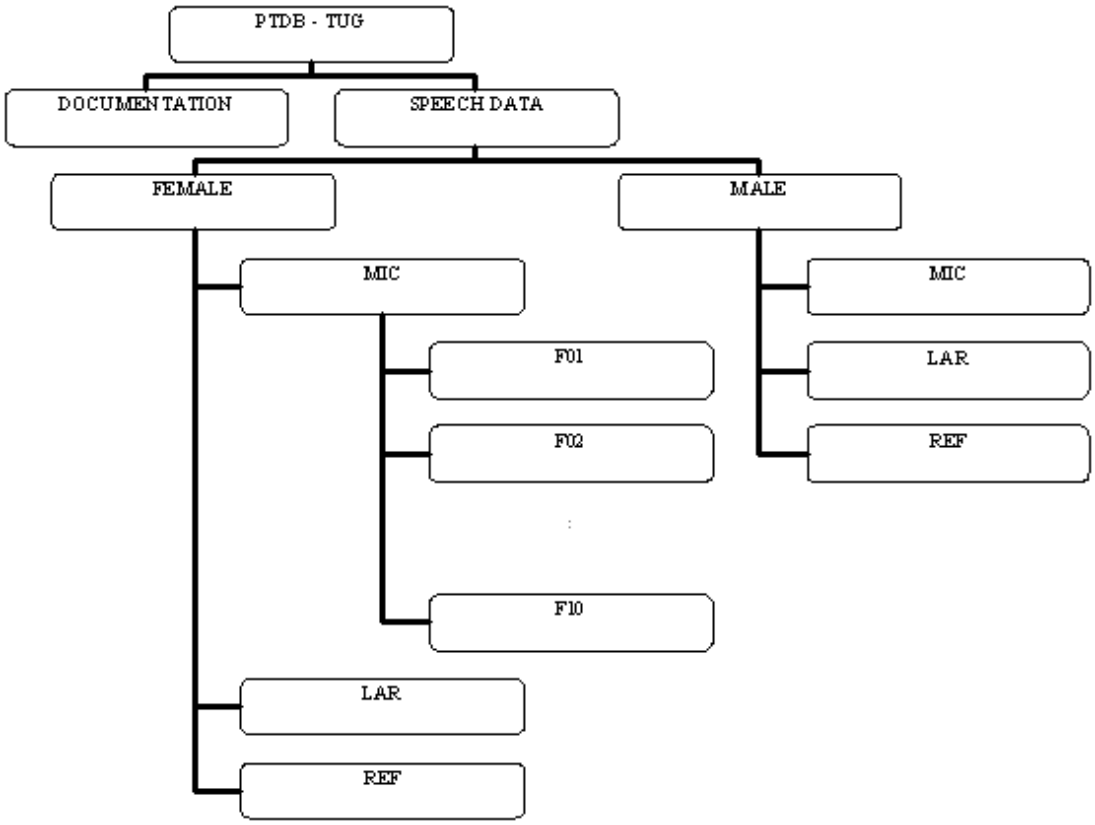

Figure 5.1: Database structure

According to figure 5.1 the PTDB-TUG consists of two subdirectories containing the documentation and the speech data. The speech signal files are separated into female data and male data in the first place and into microphone signals, laryngograph signals and reference signals in the second place. In each of these three directories one can find folders labeled according to the speaker IDs which contain corresponding data.

In the documentation directory beside this PDF the following files are available: RECORDING-PROTOCOL.TXT, SPEAKER-PROFILES.TXT and TIMIT-PROMTS.TXT.

One can choose between female and male data based on the same spoken content. Both directories, FEMALE and MALE, provide three signal categories:
MIC: Signals recorded by microphone
Lar: Signals recorded by laryngograph
Ref: Extracted reference pitch trajectories

5.2 Terminology

Table 5.1 provides an overview of the database's terminology.

Recording ID	`[Data category]_[ Speaker ID]_[Sentence ID]`
Data category	mic ... microphone signal lar ... laryngograph signal ref ... reference pitch trajectory
Speaker ID	`[Sex][Speaker number]`
Sex	F ... female, M ... male
Sentence ID	adopted from the TIMIT prompts according to Table 4.1

Table 5.1: Terminology of the PTDB-TUG

For instance `mic_F04_sa2.wav` is a WAV file providing sentence sa2 read by speaker F04 and recorded by microphone. Speaker F04 is the female speaker number four.

6 Data Recording

6.1 Acoustical Environment

In order to produce high quality signals in a defined acoustical environment with the possibility to control and modify this process immediately, the appropriate setup for this speech corpus production had to be a supervised on-site recording in a recording studio. Consequently, all recordings were done in the recording studio at the Institute of Broadband Communications at Graz University of Technology. Figure 5.1 shows the recording setup: Both, the speaker and the supervisor were sitting in the recording room and were looking at their own screen. The supervisor controlled and monitored the recording procedure with the help of the recording software SpeechRecorder [3] and headphones. The speaker was equipped with the headset microphone and the neck band with the laryngograph electrodes and had to read the displayed sentence. To reduce the background noise from the recording laptop, the supervisor position was separated by an absorbing wall from the speaker.

Figure 6.1: Recording Setup

6.2 Recording procedure

Special attention was paid on the placement and the distance of the headset as well as on the position of the laryngograph electrodes. The headset microphone had to be at a distance of about one or two cm from the speakers corner of the mouth. The right position for the electrodes of the laryngograph is on either side of the larynx. Before starting the received signals were adjusted and tested by means of some extra sentences. No special instructions were given to the probands except for having to read the displayed sentences. The recordings were made with the help of the particular speech recording program SpeechRecorder [3] and were carried out sentence by sentence, so that repetitions in case of reading mistakes or technical problems such as signal clipping could be done easily. According to the recording phases in [3] each utterance was recorded with a predelay of 2000 ms to give the speaker a certain time to get prepared and a postdelay of 500 ms to avoid signal truncation due to stopping the recording too early. The whole recording session took about an hour and contained breaks if desired by the participants.

6.3 Technical Setup

For this recording task an IBM laptop, type 2366, equipped with the program SpeechRecorder [3] as well as the firewire recording interface Presonus Firebox was used. All microphone signals were recorded by means of an AKG HC 577 L condenser headset microphone with omni-directional pickup pattern. Additionally, the vocal folds vibration was detected by a so-called Portable Laryngograph®.

7 Post-processing

7.1 Microphone and Laryngograph Signals

The provided microphone and laryngograph signals in this database were digitized at 48 kHz and 16 bit resolution. The two signal types were recorded in stereo wav-files. The left channel was used for the microphone, the right channel for the laryngograph. Later on, the channels were extracted into mono wav-files and renamed. No further post-processing (cutting, filtering, ...) was carried out on this data.

7.2 Reference Signals

The reference pitch trajectories, which are provided as ground truth data, were extracted out of the laryngograph waveforms. In general, a laryngograph signal recorded during voiced speech shows a quasi-periodic shape that represents the vocal folds vibration. Additionally, a lower frequency component is superimposed on this shape, which is mainly caused by larynx movement. Before pitch extraction can be carried out, this part has to be removed by a high pass filter to reduce pitch candidates, that deviate from the true pitch trajectory (outliers). Filtering was carried out in Matlab by applying a linear phase Kaiser filter with parameters $\beta = 0{,}5$ and n = 2400 to the rough laryngograph signals. For each gender group one specific cut-off frequency f_c was used. For the female speaker signals the cut-off frequency was $f_c = 25Hz$, for the male $f_c = 15Hz$. Finally, the RAPT algorithm [6] was run on the filtered laryngograph signals to extract the pitch. The RAPT algorithm is implemented in Wavesurfer [8] and can be applied to multiple wave-files in batch mode using a scripting language called the snack sound toolkit [9]. The output of RAPT provides 4 measures per time frame:

1st column: the pitch estimate [Hz]

2nd column: probability of voicing

3rd column: local root mean squared estimate (RMSE)

4rd column: value of peak normalized cross-correlation value that was detected and used to determine the pitch estimate.

The reference pitch was extracted using a 32ms analysis window with 10ms hopsize. Depending on the application, the user might want to extract the pitch with different settings for the analysis window. This can be done quite easily with [9] as follows

We used the following script (extractPitch.tcl) for pitch extraction:

```
1  #!/bin/sh
2  # the next line restarts using wish \
3  exec wish8.4 "$0" "$@"
4
5  package require snack
6
7  snack::sound s
8
9  foreach file $argv {
10  s read $file
11
12   set fd [open [file rootname $file].f0 w]
13   puts $fd [join [s pitch -method esps -windowlength 0.032 -framelength 0.01] \n]
14   close $fd
15 }
16
17 exit
```

Listing 7.1: extractPitch.tcl: tcl script for pitch extraction using the RAPT algorithm.

Assuming you have installed tcl (tool command language) (see [9] for details), you can then apply RAPT from your shell on multiple wave files:

```
>> tclsh extractPitch.tcl *.wav
```

where the script (see listing 7.1) as well as all wave files are assumed to be located in the current working directory. The user can set a different window length or hopsize in line 13.

8 Conclusion

This report introduced the pitch tracking database PTDB-TUG produced at Graz University of Technology. The database consists of microphone and laryngograph recordings of 10 female and 10 male speakers, who had to read 2342 phonetically rich sentences. The spoken language of this corpus is English. In addition to the recordings, the corresponding ground truth pitch signals are provided. This report also describes the production process including the recording and the subsequent signal processing, concerning mainly the reference pitch trajectories. The PTDB-TUG is the first of its kind that contains a large number of subjects speaking a great number of different utterances. For research purposes, the database can be downloaded from the website of the SPSC Laboratory at Graz University of Technology.

Bibliography

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

[2] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett and N.L. Dahlgren, "The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM," NTIS, order number PB01-100354, 1993, now available from LDC.

[3] C. Draxler, "Speech recorder quick start and user manual," Institute of Phonetics and Speech Processing, University of Munich, Tech. Rep. www.speechrecorder.org, 2011.

[4] M. Wohlmayr, M. Stark, and F. Pernkopf, "A probabilistic interaction model for multipitch tracking with factorial hiddenMarkov models," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 799-810, 2011.

[5] M. Wohlmayr, R. Peharz, and F. Pernkopf, "Efficient implementation of probabilistic multipitch tracking," in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2011.

[6] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," Speech Coding and Synthesis, W. B. Kleijn and K. K. Paliwal, Eds., pp. 495-518, 1995.

[7] G. Pirker, M. Wohlmayr, S. Petrik and F. Pernkopf, "A Pitch Tracking Corpus with Evaluation on Multipitch Tracking Scenario," in Interspeech, 2011.

[8] http://www.speech.kth.se/wavesurfer/

[9] http://www.speech.kth.se/snack/