# A Pitch Tracking Corpus with Evaluation on Multipitch Tracking Scenario

*Gregor Pirker, Michael Wohlmayr, Stefan Petrik, Franz Pernkopf*

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Graz, Austria
gregor.pirker@student.tugraz.at, {michael.wohlmayr, stefan.petrik, pernkopf}@tugraz.at

## Abstract

In this paper, we introduce a novel pitch tracking database (PTDB) including *ground truth* signals obtained from a laryngograph. The database, referenced as PTDB-TUG, consists of 2342 *phonetically rich* sentences taken from the TIMIT corpus. Each sentence was at least recorded once by a male and a female native speaker. In total, the database contains 4720 recordings from 10 male and 10 female speakers. Furthermore, we evaluated two multipitch tracking systems on a subset of speakers to provide a benchmark for further research activities. The database can be downloaded at http://www.spsc.tugraz.at/tools.
**Index Terms**: pitch tracking database, multipitch tracking

## 1. Introduction

The estimation of pitch is important for a number of speech applications such as prosody analysis, speech coding, speaker identification, speech enhancement, or speech recognition – particularly for tonal languages. Over the last decades, many different algorithms for pitch tracking have been proposed, where the best performing algorithms for single pitch estimation are YIN [1] and RAPT [2]. The term *pitch* is usually used in psychoacoustics to describe a perceptual quality. The *fundamental frequency* which denotes the inverse of the smallest period of a quasi-periodic speech signal often corresponds to the perceived pitch, however perceptual phenomena, e.g. pitch doubling, can only be explained sufficiently in the field of psychoacoustics. Since in most of the literature the term pitch is used for the fundamental frequency, we also maintain this tradition within this article.

There are two databases available, namely the Mocha-TIMIT [3] and the Keele [4] corpus, which also include laryngograph recordings of the spoken utterances. The Mocha-TIMIT data consists of 460 English sentences from a male and a female speaker sampled at 16kHz. The Keele corpus consists of five male and five female speakers reading the same phonetically balanced text of about 40 seconds duration.

In [5], we developed a fully probabilistic model for multipitch tracking which includes a speaker interaction model and a factorial hidden Markov model (FHMM). Details about the model are provided in Section 3. All parameters of the model can be learned from data either in a speaker dependent (SD) or a speaker independent (SI) fashion. SD modeling means that in case of known speaker identities and available training material we can learn the models specifically for those speakers. Learning this FHMM model using SD or SI data in a statistically robust manner requires both, sufficiently available training data and recordings from many individuals. Both existing databases are lacking in these aspects, i.e. Mocha-TIMIT provides data for only two speakers, whereas the Keele corpus is restricted in the amount of data provided for each speaker.

In this paper, we introduce a new corpus, called PTDB-TUG, composed of 2342 phonetically rich English sentences taken from the TIMIT corpus. Each utterance was recorded at least once by a male and a female speaker. The database contains 4720 audio and laryngograph recordings. The reference pitch values were extracted from the laryngograph recordings using the RAPT method [2]. Before *ground truth* extraction a Kaiser high pass filter is applied to the raw laryngograph signals to remove artefacts caused by larynx movement. Additionally, we provide a comparison of two state-of-the-art multipitch tracking algorithms using a subset of speakers of PTDB-TUG. The algorithms are our probabilistic approach based on FHMMs and speaker interaction models [5] and the method of Wu et al. [6] which is based on the auditory model of pitch perception using correlograms. The SD optimized FHMM-based approach significantly outperforms the correlogram-based method. Speaker independent FHMMs are only slightly better compared to [6]. Similar observations have been reported for other datasets in [5]. In particular, the main difference between both methods is that SD information can be easily incorporated which improves the speaker assignment of the pitch trajectories.

The paper is organized as follows: Section 2 outlines the specification of the PTDB-TUG corpus including the recording setup and post-processing. In Section 3, we compare two multipitch tracking approaches on a subset of speakers from PTDB-TUG. Section 4 concludes the paper.

## 2. PTDB-TUG Corpus

Due to the above mentioned limitations of Mocha-TIMIT and the Keele corpus, we decided to record a larger dataset including $\sim$230 utterances from each of the 20 individuals. In the following, we outline the technical specifications, the recording setup, the post-processing, and availability information. Further details can be found in [7].

### 2.1. Specifications

The PTDB-TUG contains the audio recordings and laryngograph signals of 20 English native speakers – 10 female and 10 male speakers – as well as the extracted pitch tracks as a reference. The text material consists of 2342 phonetically rich sentences, which are taken from the TIMIT corpus [8]. Each sentence was read at least once by both a female and a male

This work was supported by the Austrian Science Fund (Project number P22488-N23) and (Project number S10604-N13). Additionally, parts of this research were carried out in the context of AAP-COMET, a joint project of Graz University of Technology, Philips Speech Processing, AKG Acoustics, and ATRONIC Austria.

speaker. In total this database consists of 4720 recorded utterances. Details about the speakers (age, mother tongue, home country) are provided in [7].

The TIMIT corpus consists of two dialect sentences (labeled as *sa*), 450 phonetically-compact sentences (labeled as *sx*), and 1890 phonetically-diverse sentences (labeled as *si*). Table 1 illustrates the distribution of these sentences among speakers in PTDB-TUG. The two *sa* sentences were read by all 20 speakers. Additionally, each speaker read 45 of the *sx* sentences and 189 of the *si* sentences. Each sentence was spoken by at least one female and one male speaker.

| Sentence type | *sa* | *sx* | *si* | Total |
|---|---|---|---|---|
| # sentences | 2 | 450 | 1890 | 2342 |
| # speakers / sentence | 20 | 1 M + 1 F | 1 M + 1 F | |
| Total # utterances | 40 | 900 | 3780 | 4720 |
| # sentences / speaker | 2 | 45 | 189 | 236 |

Table 1: Distribution of *sa*, *sx*, and *si* sentences among the male (M) and female (F) speakers for the PTDB-TUG corpus.

### 2.2. Recording Setup

In order to produce high quality recordings in a defined acoustical environment with the possibility to control and modify this process immediately, the appropriate setup for this speech corpus was a supervised on-site recording in a recording studio. The studio room was equipped with two seats and screens for the supervisor and the speaker separated by an absorbing wall to reduce the background noise. The supervisor controlled and monitored the recording procedure with the help of the recording software *SpeechRecorder* [9] and headphones. The uttered sentences were recorded with a headset. Additionally, the speaker had to carry a neck band with the laryngograph electrodes. The test persons had to speak the sentences displayed on the screen. Both, microphone signals and laryngograph signals, were sampled at 48 kHz with 16 bit resolution.

### 2.3. Post-processing

The database contains the recorded signals from the microphone and the laryngograph. Furthermore, we provide extracted reference pitch trajectories from the laryngograph signals. However, users of the database may also derive their own reference signals from the raw laryngograph data. The laryngograph signals exhibit high frequency oscillations which are directly linked to the vocal folds vibrations. However, low frequency disturbances are present too which are mainly caused by the vertical movement of the larynx. We decided to filter the raw laryngograph signals with a high pass filter in order to suppress these artifacts during extraction of the reference signals using the RAPT method [2]. Empirically we observed that a Kaiser filter with a cut-off frequency of 50Hz and 30Hz for females and males, respectively, is sufficient to remove the low frequency components from the larynx movement.

### 2.4. Corpus Validation and Availability

The database was validated by an external validator who was not involved in the specification and recording process. The validation included inspection of the signal files with respect to format, sound length, clipping, and DC offset. Furthermore, a small set of randomly chosen transcriptions ( 1%) was manually compared with their corresponding signal files for accuracy and completeness. The PDTB-TUG database and documentation is available for download at http://www.spsc.tugraz.at/tools.

## 3. Multipitch Tracking

Multipitch tracking estimates the pitch of multiple concurrent speakers. A potential application is single channel speech separation [5] or automatic transcription of music. We evaluate two multipitch tracking algorithms on a subset of speakers provided in PTDB-TUG. In the following, we shortly summarize both tracking approaches.

### 3.1. Correlogram-based Algorithm

This method is based on the unitary model of pitch perception [10]. Wu et al. [6] introduced several improvements which results in a probabilistic representation of the periodicities in the signal. First, the input signal is decomposed into 128 subbands using a gammatone filterbank, and the amplitude envelope is extracted for high-frequency channels (center frequency above 800 Hz). The normalized autocorrelation function is determined framewise on each channel. Furthermore, channels with unreliable periodicity because of noise are removed. Low-frequency channels are selected in cases where the maximum peak at nonzero lags is above a threshold. High-frequency channels are selected if the periodicity information is consistent with the autocorrelation at a larger time frame. Additionally, a peak selection routine is applied. Finally, the set of peaks selected from all channels serves as basis for a probabilistic representation of zero, one or two pitch periodicity values at each time frame. Basically, a likelihood of pitch periodicities under the given observation for the hypothesis of one and two pitch values is determined. Subsequently, these likelihoods are modeled by a hidden Markov model which leads to semi-continuous pitch trajectories. This model is able to provide an excellent performance in terms of pitch estimation accuracy, however, the assignment of speakers to the pitch estimates is inaccurate, i.e. it is not possible to correctly link each pitch estimate to its source speaker.

### 3.2. FHMM-based Algorithm

In [5], we proposed to use a statistical model using speech mixture spectograms as observations. This approach is quite different from the auditory-based methods as proposed above [6]. We do not require any heuristics such as peak or channel selection. These are implicitly included in the statistical model. However, we do need to parameterize this model, i.e. we have to learn the parameters using training data. The specification of the model parameters can be done either in SD or SI manner. Our multipitch tracking method consists of the following modules:

**1)** Gaussian mixture models (GMMs) are used to model the spectrogram of each single speaker. The number of Gaussian components is determined using the minimum description length (MDL) criterion. GMMs can be either trained on a large set of different speakers, or if prior knowledge about the speaker identities is available, the GMMs can be optimized for single speaker data. This results in SI or SD models, respectively. In Section 3.3, we provide results for both, SI and SD models. SD models are usually more accurate and offer the advantage of correct assignment of pitch trajectories to the corresponding speakers. This is important for e.g. single-channel speech separation.

**2)** We use the MIXMAX speaker interaction model [11] to ob-

tain a probabilistic representation of the observed speech mixture of both speakers. The fundamental assumption of the MIXMAX model is that speech is sparse in time-frequency representations, i.e. each particular time-frequency bin of a speech mixture spectrogram is dominated by one speaker. Hence, the log-spectrum of two speakers can be approximated by the elementwise maximum of two single speaker log-spectra. This MIXMAX model is related to the concept of binary masks in computational auditory scene analysis (CASA) [12]. In [5], an alternative linear interaction model is also considered.

**3)** The statistical speaker interaction model for the speech mixture is used within the framework of FHMMs [13]. FHMMs enable tracking the pitch trajectories of both speakers. Each hidden Markov process of the FHMM models the pitch trajectory of a particular speaker, where the available observations are considered as a joint effect of all individual Markov processes. The explicit factorial nature among the various Markov chains allows to use more efficient inference algorithms compared to an equivalent HMM. Here, we use exact inference mechanisms to extract the pitch trajectories.

Recently, we significantly improved the computational efficiency of FHMM-based multipitch tracking [14]. We show that the tracking performance is almost unaffected when discarding up to 99.5% of the smallest likelihood values in the observation model. Following this observation, we proposed two methods to efficiently find the largest likelihood entries. This results in significant time savings for likelihood computation as well as for tracking by making use of sparse likelihood matrices.

### 3.3. Results on PTDB-TUG

We evaluate both algorithms, i.e. the correlogram-based and the FHMM-based approaches (abbreviated as CORR and FHMM, respectively), in terms of tracking performance. We selected two male speakers (M04 and M10) and two female speakers (F01, F07) from the set of 20 speakers. For each speaker we select 224 utterances from the *sx* and *si* sentences to train the speaker dependent FHMM model, whereas the remaining 10 sentences were used for performance evaluation. Since speech mixtures for each speaker pair (male-male, male-female, and female-female) and utterance are produced, we have in total 600 test mixtures (i.e. 100 mixtures for each of the six speaker pairs). The speakers are mixed at 0dB signal-to-signal ratio. The reference pitch trajectories are extracted as described in Section 2.3. Speaker independent FHMM models are optimized on all sentences of the following speakers: M01, M05, M06, M07, M08, M09 F02, F04, F05, F06, F08 and F10. The test mixtures remain the same as for the speaker dependent case. Furthermore, long silence intervals at the beginning and the end of the utterances have been removed, i.e. the utterances are cut at 70ms before the first and 70ms after the last occurrence of pitch values.

The correlogram-based approach requires as input the speech mixture in time-domain, whereas the FHMM-based method relies on the log-spectrogram $\mathbf{y}$ of the speech mixture. First, we resample the signals of the database to a sampling rate of 16kHz and compute a log-spectrogram using a 1024-sample FFT of 32ms Hamming windowed segments and a 10ms step size. Each observation vector is composed as magnitude of the spectral bins 2-65, which corresponds to a frequency range up to 1000 Hz. During tracking we use the likelihood pruning mechanisms as proposed in [14].

We measure the performance of both algorithms in terms of the error measure $\bar{E}_{Total}$ which also accounts for speaker assignment errors. This measure slightly deviates from the error measure proposed in [6]. Each of the two estimated pitch trajectories $\tilde{f}_0^1[t]$ and $\tilde{f}_0^2[t]$ is assigned to the reference trajectory provided in the database, i.e. $f_0^1[t]$ or $f_0^2[t]$. Basically, there are two assignments possible, either $\left(\tilde{f}_0^1[t] \rightarrow f_0^1[t]; \tilde{f}_0^2[t] \rightarrow f_0^2[t]\right)$ or $\left(\tilde{f}_0^1[t] \rightarrow f_0^2[t]; \tilde{f}_0^2[t] \rightarrow f_0^1[t]\right)$. We select the assignment with the smaller quadratic error over all time instances. Note that this assignment is performed globally and not locally for each individual time frame.

The total error is composed as $\bar{E}_{Total} = E_{01}+E_{02}+E_{10}+E_{12}+E_{20}+E_{21}+\bar{E}_{Gross}+\bar{E}_{Fine}+\bar{E}_{Perm}$, where $E_{ij}$ denotes the percentage of time frames, where $i \in \{0,1,2\}$ pitch points are missclassified as $j \in \{0,1,2\}$ pitch points, $i \neq j$. $\bar{E}_{Perm}$ measures the percentage of frames, where the voicing decision is correct, but the pitch values are not assigned to the correct speakers. $\bar{E}_{Gross}$ is the percentage of frames with correct voicing decision and no permutation error, where at least one detected pitch value deviates more than $20\%$ from the reference. The frequency deviation is determined as

$$\Delta f_0^i[t] = \frac{|\tilde{f}_0^i[t] - f_0^i[t]|}{f_0^i[t]}.$$

$\bar{E}_{Fine}$ is composed of $\bar{E}_{Fine}^1+\bar{E}_{Fine}^2$, where $\bar{E}_{Fine}^i$ represents the frequency deviation for speaker $i$ in percent, averaged over frames where no voicing, no gross and no permutation errors are present.

Figure 1 shows the total error $\bar{E}_{Total}$ and its standard deviation for each speaker pair on all 600 test speech mixtures for CORR, speaker independent FHMMs, and speaker dependent FHMMs, respectively. The $x$-axis represents the corresponding speaker pairs.

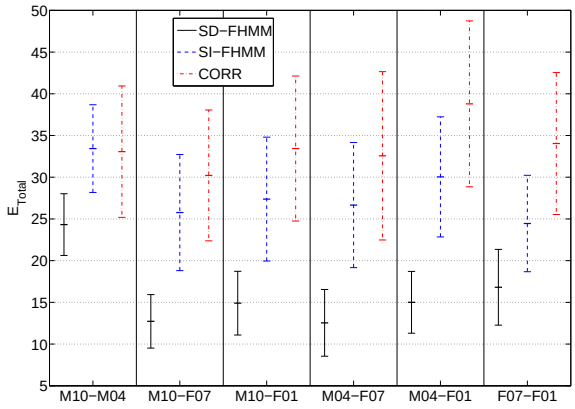


Figure 1: Total error for each speaker pair of 600 test speech mixtures using correlogram-based multipitch tracking (CORR) and speaker dependent and independent FHMM-based multipitch tracking (SD-FHMM and SI-FHMM, respectively).

SI FHMMs slightly outperform CORR on most speaker pairs. Speaker dependent FHMMs have a significantly lower $\bar{E}_{Total}$ compared to CORR. Looking at individual speech mixture examples (one example is shown in Figure 2) it turns out that the main factor for the superior FHMM-based multipitch tracking performance is an improved speaker assignment for the

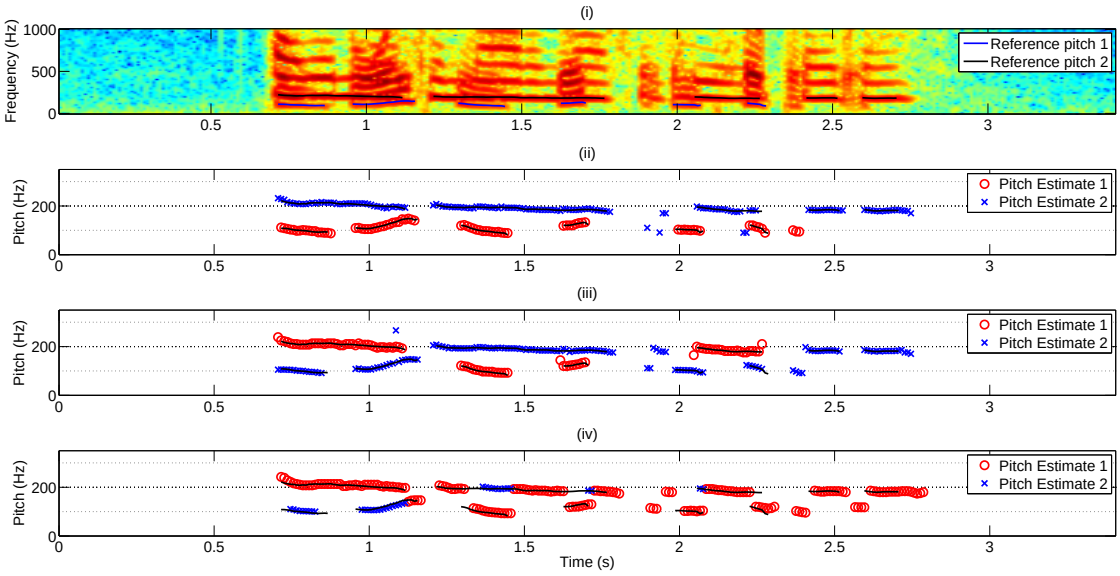


Figure 2: Tracking example for a speech mixture; (i) Log-spectrogram of speech mixture with reference pitch trajectories; (ii) Estimated pitch trajectories with reference using SD-FHMMs ($\bar{E}_{Total} = 9.9677$); (iii) Estimated pitch trajectories with reference using SI-FHMMs ($\bar{E}_{Total} = 24.0397$); (iv) Estimated pitch trajectories with reference using CORR ($\bar{E}_{Total} = 28.5585$).

respective pitch trajectories. This was already observed for different datasets in [5]. One advantage of CORR is that it can be applied to a test speech mixture without any training phase. The FHMM-based models require parameter learning.

## 4. Conclusions

We introduced a new pitch tracking corpus containing 4720 audio and laryngograph recordings from 10 male and 10 female speakers. Existing databases either do not have a sufficient variety of speakers or do not contain enough recorded material per speaker for serious training. Furthermore, we provide reference pitch tracks extracted from the laryngograph recordings.

Additionally, we apply two multipitch tracking approaches on a subset of speakers from PTDB-TUG. The probabilistic approach based on FHMMs and speaker interaction models can be learned on speaker dependent (SD) and speaker independent (SI) data whereas the correlogram-based approach does not require any training. The SI FHMM approach is slightly better that the correlogram-based method. The SD optimized FHMM-based method significantly outperforms both approaches. Similar observations for other datasets are reported in [5]. The main benefit of the SD-FHMM-based method is that it improves the speaker assignment of the pitch trajectories.

In future, we aim to adapt speaker independent FHMM-based models to speaker specific characteristics during multipitch tracking in an expectation-maximization-like manner.

## 5. References


[1] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[2] 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.

[3] A. Wrench, "A multichannel/multispeaker articulatory database for continuous speech recognition research," *Phonus*, vol. 5, pp. 3–17, 2000.

[4] F. Plante, G. Meyer, and A. Ainsworth, "A pitch extraction reference database," in *Eurospeech*, 1995, pp. 837–840.

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

[6] M. Wu, D. Wang, and G. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 11, no. 3, pp. 229–241, 2003.

[7] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "Database for multi-pitch tracking," Graz University of Technology, Signal Processing and Speech Communication Laboratory, Tech. Rep. www.spsc.tugraz.at/tools, 2011.

[8] L. Lamel, R. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Proceedings of the DARPA Speech Recognition Workshop, Report No. SAIC-86/1546*, 1986.

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

[10] R. Meddis and L. O`Mard, "A unitary model of pitch perception," *J. Acoust. Soc. Am.*, vol. 102, no. 3, pp. 1811–1820, 1997.

[11] A. Nadas, D. Nahamoo, and M. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 10, pp. 1495–1503, 1989.

[12] D. Wang and G. Brown (Eds.), *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. IEEE Press/Wiley-Interscience, 2006.

[13] Z. Ghahramani and M. Jordan, "Factorial hidden Markov models," *Machine Learning*, vol. 29, no. 2-3, pp. 245–273, 1997.

[14] M. Wohlmayr, R. Peharz, and F. Pernkopf, "Efficient implementation of probabilistic multi-pitch tracking," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. –.