

# Automatic Reconstruction of Utterance Boundaries Time Marks in Speech Database Re-grabbed from DAT Recorder

Hynek Bořil

Czech Technical University in Prague, Faculty of Electrical Engineering

Technická 2, 166 27 Prague, Czech Republic

Email: borilh@gmail.com

**Abstract.** *In this paper, an algorithm performing automatic reconstruction of utterance boundaries time marks in speech database re-grabbed from DAT recorder is presented.*

*Originally, the database was grabbed from DAT and, after down-sampling, processed at 16 kHz. Utterance boundaries were manually found, each utterance was stored to a separate file and orthographic and phonetic transcriptions were performed.*

*Recently, a requirement to re-grab and process the database at 48 kHz has appeared. Since positions of utterance boundaries were known for 16 kHz, it was reasonable to use this information for the 48 kHz processing. Unfortunately, re-grabbed sessions displayed certain length changes compared to the originally grabbed data.*

*Presented algorithm finds session boundaries, matches session lengths and calculates actual utterance boundaries positions. Finally, matching accuracy is automatically checked.*

**Keywords:** speech database processing, preannotation, asynchronous sample rate conversion.

## 1. Introduction

Concerning the data storage process, generally, there are two common approaches to speech database recording. First possibility is to store the recorded session by utterances, i.e. every utterance is stored in a separate file. If an item is wrongly uttered, speaker may be asked to repeat it again. In such a case, no preprocessing before annotation is needed. This approach, used for example for recording of the Czech SPEECON and CLSD databases [1, 2], requires permanent attention of the operator during the whole recording. In some

recording environments, the operator may be occupied by another activities, e.g. by driving a car. In this case it appears reasonable to record the session as a whole. Speaker is usually asked to correct eventual mistakes by himself/herself without initiation of the operator, e.g. in CZKCC car database [3]. Such sessions have to be preannotated, i.e. boundaries of all utterances are found, every utterance is stored into a separate file and wrong utterances are omitted.

In this paper, an algorithm performing automatic reconstruction of utterance boundaries time marks in speech database re-grabbed from DAT recorder is presented.

The database was recorded to DAT tapes, grabbed digitally to hard disc, down-sampled from 48 kHz to 16 kHz and processed. The 48 kHz data were then archived only on the DAT tapes.

Recently, a need to grab the database from DAT again and process it at the sampling frequency 48 kHz has risen. Since information about positions of utterance boundaries were known for the 16 kHz version, it seemed to be reasonable to use it for automatic preannotation of the 48 kHz data. But in spite of the fact that the data were grabbed from DAT using digital input of the soundcard, it was found, that the previously and recently grabbed session recordings do not just differ in the beginning time (as the tape was not played from the exactly same position), but also in duration of every utterance.

Presented algorithm finds session approximate boundaries, stretches sessions to the same length, finds session boundaries precisely, calculates new time marks from the old ones and checks whether the boundaries were determined correctly.

Finally, correctness of the automatic boundaries matching is checked for chosen samples by hearing tests.

## 2. Source Data

The originally grabbed data are stored as separate 16 KHz raw audio files for each utterance. To each session, a neg file comprising orthographic and phonetic transcriptions [4] and time positions of all session utterances are assigned. The re-grabbed data are stored as 48 kHz session files, i.e. every session is stored as a whole to a separate raw audio file. The goal is to find utterance boundaries in the session audio files and save utterances to separate files.

## 3. Session Length Variations

Although the audio data were stored on the DAT tapes and soundcard digital input S/PDIF was used for grabbing, lengths of the originally and actually grabbed sessions did not match, even if the first utterances of the equivalent sessions were aligned manually.

When transferring the contents of a DAT tape using S/PDIF, a bit-for-bit copy is obtained if the soundcard can act as a slave and follows the clock present in the arriving signal. If the soundcard does not support the slave mode, it records a digital signal, but the clocks are not locked together. Several soundcards provide fixed-frequency internal processing at 48 kHz (e.g. SoundBlaster Live!) and thus cannot be slaved to the external digital signal. In this case, asynchronous sample rate conversion (ASRC) is used. With ASRC, the signal is digitally resampled to 48 kHz. Distortion of the digitally resampled signal is still far lower than of the analogue signal that has been passed through an A-D converter [5].

In our case, soundcards employing ASRC were used. Since different DAT recorder and soundcard were used for grabbing and re-grabbing, different clocking affected originally and recently grabbed audio data.

## 4. Seek, Shrink & Mark Algorithm

By manual matching of a couple of sessions it was found, that due to ASRC, newly grabbed sessions are systematically contracted in length in comparison to the originally grabbed ones. An example of

session length variations is shown in Tab. 1, where ‘Ses. #’ refers to session number (label),  $T_{1st, last, neg, new}$  time positions of first, last utterance in the originally (neg file values) or newly grabbed session respectively.

Ses. #	$T_{1st\_neg}$ (s)	$T_{1st\_new}$ (s)	$T_{last\_neg}$ (s)	$T_{last\_new}$ (s)	$R_{fs}$ (-)
00136	3.94	6.33	1603.08	1581.94	0.985
04205	8.19	12.92	1638.83	1619.67	0.985
05578	8.31	7.79	1940.55	1908.82	0.984

Table 1. Session boundaries localization

$R_{fs}$  denotes a ratio of original session shrinking required to match the actual session length. The shrinking is reached by original utterance re-sampling at frequency

$$f_{s\_new} = R_{fs} \cdot f_s \quad (1)$$

### 4.1 Utterances Seeking

For automatic seeking of utterance boundaries in newly grabbed session, maximum of cross-correlation function [6] was detected

$$R_{U,S}(k) = \sum_{n=0}^{N_u-1} s_s(n+k) \cdot s_u(n), \quad (2)$$

$$k_{start} = \underset{k}{\operatorname{argmax}} R_{U,S}(k), \quad k = 0, 1, \dots, N-1, \quad (3)$$

where  $s_s$  is re-grabbed session signal,  $s_u$  originally grabbed utterance signal,  $N_u$  length of the utterance signal in samples,  $k$  cross-correlation shift and  $k_{start}$  detected starting position of the utterance  $s_u$  in the session signal  $s_s$ .  $N$  denotes maximum window length where the utterance start is expected to appear.

To perform the utterance seeking, both new session files and old utterance files had to be resampled to the same frequency. Since cross-correlation evaluation at 16 kHz would be computationally expensive, all signals were down-sampled to 2 kHz as the tests displayed that the accuracy remained suitable for the boundary matching. Since the shrink ratio  $R_{fs}$  was found to be almost similar for all the sessions (see Tab. 1), the old utterance files were initially resampled to

$f_s = 1971$  Hz (empirically determined as a quasi-optimal shrink sample frequency). For the purposes of the boundary matching, only first and last utterance positions in the session were determined, see Fig. 1.

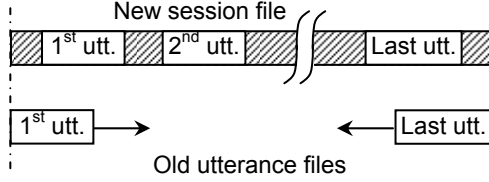


Figure 1. Session boundaries localization

Since the initial shrink ratio was just approximate, the session boundaries might not be determined precisely, as cross-correlated signals could be still of slightly different length on the utterance level. Hence, the approximate shrink ratio could be determined from the detected boundaries

$$R_{fs} = \frac{T_{last\_new} - T_{1st\_new}}{T_{last\_neg} - T_{1st\_neg}}, \quad (4)$$

where  $T_{1st\_new}$  and  $T_{last\_new}$  refer to detected starting positions of the first and last utterance, see Fig. 2.

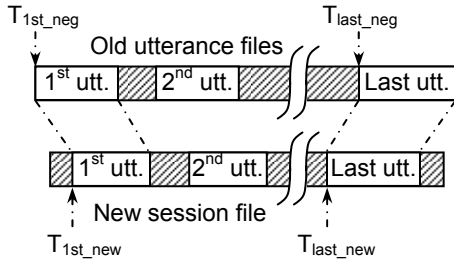


Figure 2. Shrink ratio determination

Once the approximate  $R_{fs}$  is known, old utterance files are resampled to match lengths of the new ones.

Again, positions of the first and last utterance are determined according to Eq. 2 and 3. In this case, error of the first and last utterance matching caused by approximate  $R_{fs}$  would be insignificant from the standpoint of the session length. Consequently, precise  $R_{fs}$  is determined from the actual boundaries as shown in Eq. 4 and Fig. 2.

Knowing the precise  $R_{fs}$  and  $T_{1st\_new}$ , new time marks can be calculated from the old neg time marks.

## 4.2 New Time Marks Calculation

New time marks can be calculated from the old neg ones as

$$T_{act\_new} = (T_{act\_neg} - T_{1st\_neg}) \cdot R_{fs} + T_{1st\_new}. \quad (5)$$

## 4.3 Automatic Matching Check

Sometimes, positions of the first and last utterances may be determined wrongly due to inappropriate initial shrink or the fact that another utterance appears to have higher mutual energy with inadequate part of the session signal. For this reason, automatic check was performed for every session. Actual position of the 'middle utterance' (utterance chosen from the middle of the session) was determined from the cross-correlation and compared to the corresponding time marks. If the values satisfied condition

$$|T_{calc} - T_{corr}| < \Delta T_{max}, \quad (6)$$

where  $T_{calc}$  is the calculated time mark,  $T_{corr}$  actual position determined from the cross-correlation and  $\Delta T_{max}$  maximal allowed difference, the session was considered to be processed correctly. In this case, utterance files were generated by chopping of the session file at the positions of the time marks. If the condition in Eq. 6 was not satisfied, positions of the first and last utterance were matched manually and the rest of the process was carried out as described in the previous paragraphs. Flow chart of the complete algorithm is shown in Fig. 3.

## 4.4 Hearing tests

To ensure correctness of the Seek, Shrink & Mark algorithm, hearing test were performed at the final stage. Couple of utterances distributed over the whole session were played while the expected content was displayed.

## 5. Results

The database consists of 260 sessions of typical durations from 20 to 30 minutes. Fully automatic processing (Celeron 2.2 GHz)

took about 7 minutes per one session – including final generation of utterance files.

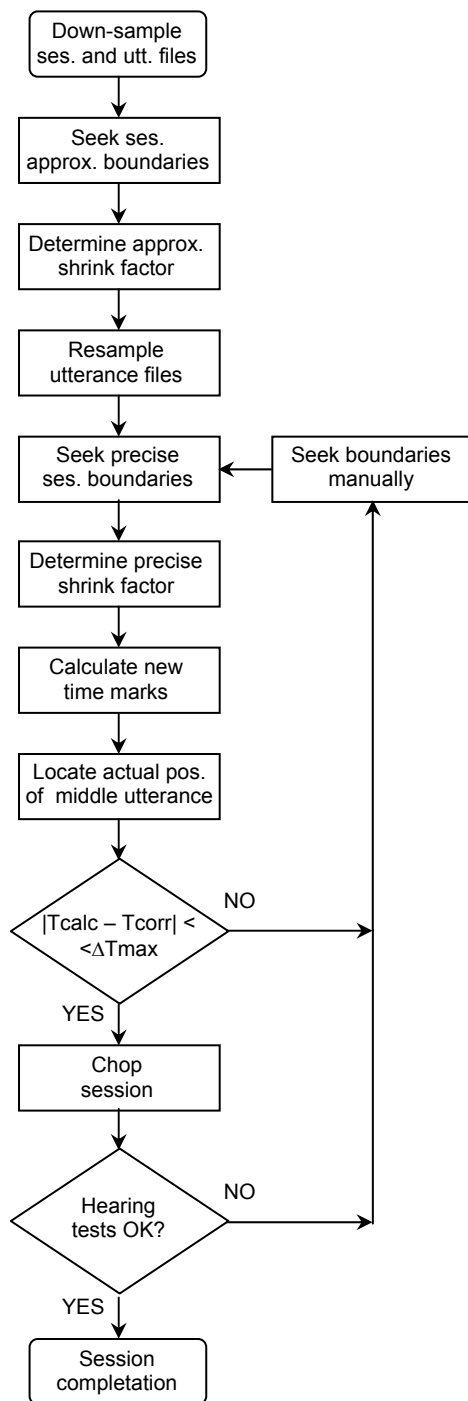


Figure 3. Seek, Shrink & Mark algorithm

Approximately 240 sessions have passed automatic matching and hearing tests without need to find session boundaries manually. About 20 sessions did not pass the automatic matching test. Most of them have passed the tests after session boundaries were found manually.

In several rare cases, there was missing part of the signal inside the re-grabbed session file, which was probably caused by omitting of the rumbled parts of the tape. In these cases, positions of the last utterance before the omitted part and the first after the omitted part had to be found in addition, the rest of the process was then also automatic. Consequently, all sessions have passed the tests.

## 6. Conclusions

In this paper, an algorithm performing automatic reconstruction of utterance boundaries in sessions re-grabbed from the DAT tapes was presented. Fully automatic processing of approximately 240 sessions has been performed. In about 20 sessions, slight manual matching was required in the beginning stage of the automatic processing.

## 7. Acknowledgements

The presented work was supported by GAČR 102/05/0278 "New Trends in Research and Application of Voice Technology", GAČR 102/03/H085 "Biological and Speech Signals Modeling", and research activity MSM 6840770014 "Research in the Area of the Prospective Information and Navigation Technologies".

## 8. References

- [1] [www.speecon.com](http://www.speecon.com)
- [2] BOŘIL, H., POLLÁK, P.: Design and Collection of Czech Lombard Speech Database, 1577-1580, *INTERSPEECH-05*, Lisbon, Portugal, 2005.
- [3] [www.temic-sds.com](http://www.temic-sds.com)
- [4] KIESSLING, A., DIEHL, F., FISCHER, V., MARASEK, K.: Specification of Databases – Specification of Annotations, Deliverable D214 of the project IST-1999-10003, 2002.
- [5] WALKER, M.: Perfect copies, all the time? Transferring digital audio using PC soundcards. Sound on Sound Magazine, [www.soundonsound.com/sos/jul99/articles/pcmusician.htm](http://www.soundonsound.com/sos/jul99/articles/pcmusician.htm), July 1999.
- [6] VEJRAŽKA, F., HRDINA, Z., Signals and Systems, ČVUT, Prague 2000, in Czech.