F0 Of Adolescent Speakers – First Results for the German Ph@ttSessionz Database

Christoph Draxler, Florian Schiel, Tania Ellbogen

BAS Bavarian Archive of Speech Signals, University of Munich, Germany
draxler@phonetik.uni-muenchen.de, schiel@phonetik.uni-muenchen.de, ellbogen@phonetik.uni-muenchen.de

Abstract

The first release of the German Ph@ttSessionz speech database contains read and spontaneous speech from 864 adolescent speakers and is the largest database of its kind for German. It was recorded via the WWW in over 40 public schools in all dialect regions of Germany. In this paper, we present a cross-sectional study of f0 measurements on this database. The study documents the profound changes in male voices at the age 13-15. Furthermore, it shows that on a perceptive mel-scale, there is little difference in the relative f0 variability for male and female speakers. A closer analysis reveals that f0 variability is dependent on the speech style and both the length and the type of the utterance. The study provides statistically reliable voice parameters of adolescent speakers for German. The results may contribute to making spoken dialog systems more robust by restricting user input to utterances with low f0 variability.

1. Introduction

There exist many studies dealing with the development of voice and speech in children and adolescents. One important measure on which many studies have focused is the fundamental frequency f0 and its variability, e.g. in (Kent, 1976), (Keating and Buhr, 1978), (Bennett, 1983), (Huber et al., 1999), (Whiteside and Hodgson, 2000), (Kovacic and Budanovac, 2002), (Whiteside et al., 2002).

Many of these studies suffer from pragmatic, but nevertheless severe limitations:

1. they analyze small numbers of speakers, usually less than 50,
2. they use carefully designed and limited speech material, e.g. sustained vowels, and
3. the original speech data used in these studies is usually not available, so that additional analyses or comparisons using the same material cannot be performed.

In their study of developmental changes of children's speech, (Lee et al., 1999) have used a speech database with more than 450 speakers. However, the speech material is again very limited, and the speech data is not available.

In forensic studies, f0 is a key factor for speaker identification. These studies often access speech databases which are either generally available, e.g. (Lindh, 2006), or specifically tailored to the task, e.g. (Nolan et al., 2006), (Hudson et al., 2007).

In this paper we demonstrate the use of a very large speech database for acoustical analyses of adolescent speakers. We focus on f0 variation, because it is a key factor in speech recognition and speaker verification.

The remainder of the paper is structured as follows: in section 2 we describe the structure and contents of the German Ph@ttSessionz speech database. Section 3 presents f0 measurements related to age and speaking style. The results are discussed in section 4 and summarized in a conclusion.

2. Ph@ttSessionz Speech Database

Ph@ttSessionz is a German speech database that in its first release contains the speech of 864 adolescent speakers between 12 and 20 years of age. The speech material is a superset of the German SpeechDat II and the RVG-1 corpus ((Burger and Schiel, 1998), (Winski, 1997)), and it consists of read and non-scripted speech.

The database is organized in recording sessions. In a given recording session, a speaker is asked to read prompts or respond to questions in a prompt sheet. A prompt sheet comprises at least 125 items (see table 1).[1]

Type	Count
isolated digits, ”zwo”	10
numbers 11-100	19
PC command phrases	12
phonetically rich sentences	30
telephone numbers with area code	3
6- and 7-digit telephone numbers	10
mobile phone keys (digits, $*$, #)	3
credit card numbers	3
PIN codes (6 digits)	3
date expressions	3
spellings	10
directory assistance names	9
time expressions	3
spontaneous responses	5
narrative speech	2

Table 1: Ph@ttSessionz database contents

2.1. Recordings

Recordings commenced in January 2005 and are still continuing. The recordings take place in public secondary schools (Gymnasium) in all dialect regions of Germany (Hollmach, 2003). They are performed via the Internet

[1] Later in the project, additional items were recorded, increasing the number of prompt items in each session to 133.

using the SpeechRecorder software (Draxler and Jänsch, 2004). During a recording session, prompt items are retrieved from a web server and signal data is immediately uploaded to the server (Draxler and Jänsch, 2006). The recordings themselves were unsupervised, i.e. after adjusting the recording level during the first five test items the recording supervisor left the room and the speaker was alone in front of the PC.[2]

The recording environments were normal class or computer rooms.

2.2. Signal quality

A standard recording equipment was used at all sites. This equipment was sent to the schools by mail and returned after the end of the recording period.

The recording equipment consists of an M-Audio Mobile Pre USB audio interface, an Audio Technica AT 3031 table microphone, and a Beyerdynamic opus 54 headset microphone. The signal quality is 22.05 kHz, 16 bit stereo, linear quantization. In general, the recorded signal is very good.

2.3. Speaker data

The following data on the speakers were collected by the recording supervisor via a web form at the beginning of every recording session:

- date of birth, sex, weight, height
- dialect region (federal state of Germany where speaker entered school)
- mother tongue of the speaker, his or her mother and father
- smoking habits, piercing in lips or tongue, dental braces

95.7% of the speakers stated German was their native language, 91.6% said their mother's native language was German, 90.4% their father's.

2.4. Transcription

The Ph@ttSessionz database is transcribed according to slightly modified SpeechDat-II transcription guidelines (Senia and van Velden, 1997). It consists of an orthographic transcript with a limited set of markers for filled pauses, articulatory noise, external noise, mispronunciations, incomprehensible speech and signal truncation.

By the end of January 2007, more than 870 speakers have been recorded, with a total of more than 110.000 utterances. More than 82% of the utterances have been transcribed, resulting in over 69 hours of speech

2.5. f0 computation

From the Ph@ttSessionz database 90829 transcribed non-empty speech files of 762 speakers from 40 recording locations were selected for analysis. The transcription of the signal contains a begin and an end boundary for the speech segment, and this segment was extracted for the analysis.

f0 was computed on the headset channel using the built-in algorithm of Praat (Boersma, 2001) with the settings given in table 2.

Parameter	value
lower frequency	75
upper frequency	400
max. candidates	15
silence threshold	0.03
voicing threshold	0.45
octave cost	0.01
octave jump-cost	0.35
voiced/unvoiced cost	0.14

Table 2: settings for Praat's f0 algorithm used in the analysis

For every audio file, the following values were computed: min f0, mean f0, max f0, and f0 standard deviation. These values were then stored in a relational database to allow efficient querying via SQL.

3. Results

3.1. f0 and age

Male and female voices differ both in mean f0 and standard deviation in the age range 13 to 19 years. f0 for female voices is consistently higher than for male voices, and so is its standard deviation.

For male voices, f0 drops dramatically between the age of 13 and 15, and is almost stable from then on. The absolute drop in f0 is 80.07 Hz (206.67 Hz - 126.60 Hz) and 4.17 Hz (126.60 Hz -122.43 Hz) respectively (black curve in figure 1).

For female voices, no such drop in f0 can be observed. f0 decreases rather steadily from age 13 to 19 by 12.04 Hz (229.91 Hz - 217.87 Hz).

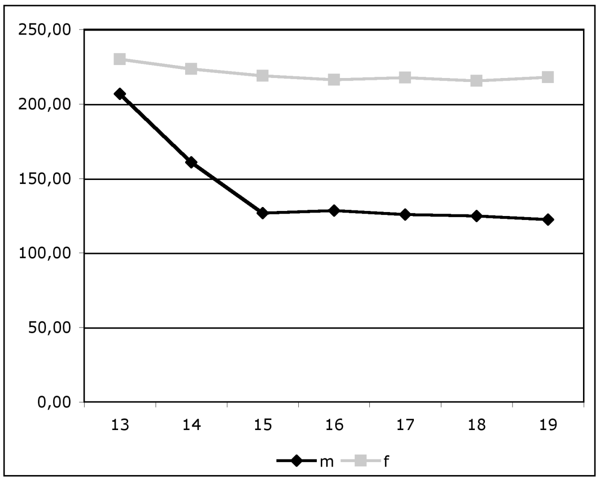

Figure 1: f0 absolute values in Hz vs. age for female (gray) and male (black) speakers

[2] In some recording sessions more than one person was present during the recordings, and this second person can be heard in the table microphone channel.

3.2. f0 variability and utterance types

The Ph@ttSessionz database contains read and non-scripted speech. The non-scripted speech is evoked by prompts such as e.g. ”Describe the clothes you are wearing today”, and ”Tell us a funny story”. The prompts for long speech production contain explicit timing instructions, e.g. ”You have one minute”. Speakers could terminate a recording any time by pressing the mouse button – which they frequently did.

For the analysis, short speech productions were only included if their duration was 2 seconds or longer, long speech productions only if their duration was 5 seconds or longer. The average length of the short and long narrative speech thus obtained was 5.5 s and 14.9 s.

To measure the variability, we calculated the f0 absolute range $f0_{abs} = f0_{max} - f0_{min}$ for every utterance type and age group. Figures 2 and 3 show two typical patterns for the range of f0 of female speakers. Note that the mean value for f0 is amost identical in both figures.

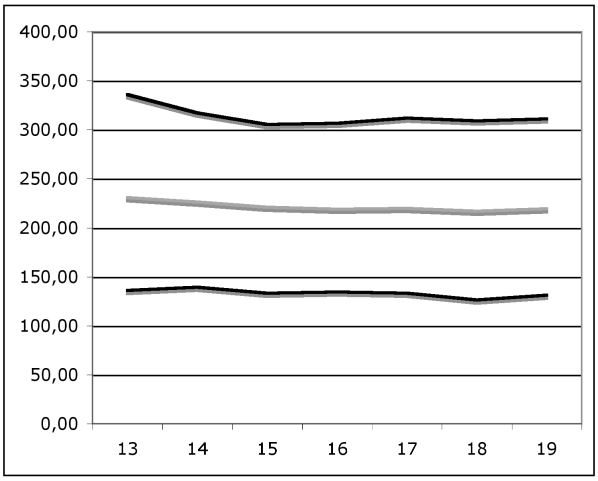

Figure 2: f0 mean and range in Hz for telephone numbers by female speakers vs. age.

In the next step, we computed f0 relative range as $f0_{rel} = f0_{max}/f0_{min}$ over the utterance types and sorted the result by average utterance duration. Two observations are worth noticing:

- $f0_{rel}$ does not depend on utterance duration alone – e.g. credit card numbers show less variability than spelled person names
- Calculated in Hertz, $f0_{rel}$ was consistently higher for female than for male speakers.

As a consequence of the second observation, we recalculated $f0_{rel}$ using the mel-scale to obtain a perception-based representation. With this scale, $f0_{rel}$ was very similar for female and male speakers, as can be seen in figure 4.

4. Discussion

From the data it can be seen that between 13 and 19 years of age male voices change dramatically and that female voices are much less affected. Our results are in line with data in

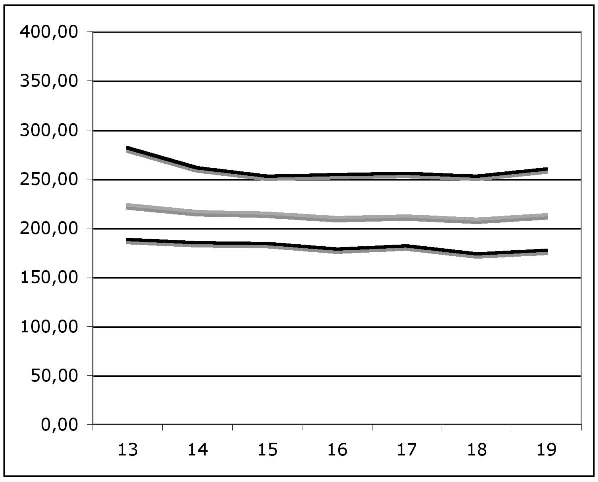

Figure 3: f0 mean and range in Hz for single digits by female speakers vs. age.

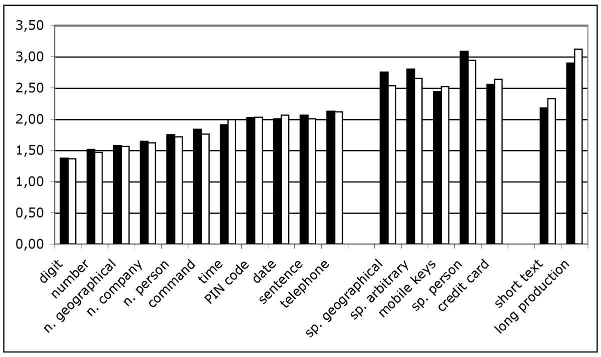

Figure 4: $f0_{rel}$ in the perceptual mel-scale vs. utterance types for female (white) and male (black) speakers. The utterance types are divided into three groups – two of read and one of spontaneous items, and sorted by increasing utterance duration from left to right within the groups.

(Lee et al., 1999), but the change occurs slightly earlier than reported in (Kent, 1976). We have also correlated speaker height and weight with f0, but like (Künzel, 1989) we did not find any correlation.

The results on f0 variability can be subdivided into three groups: group I comprises the read items with similar f0 range for male and female speakers, i.e. ’digit’ to ’telephone’ in figure 4. Group II contains read items with substantially increased variability, and group III the spontaneous items. Groups II and III also show marked differences of variability between female and male speakers.

The items in group I are short, ranging from 0.85 to 3.29 s and an average word count of 1.13 to 7.24, and have no or few linguistic structures, e.g. single digits and numbers or sentences, date expressions and telephone numbers respectively. They can be captured at a glance and be spoken naturally with very little mental workload.

Group II items are spellings and long digit sequences – their duration is between 3.78 and 7.28 s, their word count between 8.48 and 15.72. One possible explanation for their high variability is that these items are too long to be produced in one single flush. Speakers must group the utterance into chunks, increasing the mental workload. Formatting the prompts makes this chunking easier, reducing the mental workload. For example, the long credit card items with their given formatting in four groups of four digits display a lower f0 variability than the shorter spelling items which consist of unformatted character sequences separated by blank spaces. However, the f0 variability for credit card numbers is clearly higher than that of the shorter items. Spontaneous speech is characaterized by non-speech phenomena such as filled pauses, or repairs, mispronuciations, and even laughter which may contribute to increased variability. However, in the present study, these phenomena were not analyzed in detail.

5. Conclusion and outlook

The first release of the Ph@ttSessionz database is now publicly available via BAS.

With the Ph@ttSessionz database it is possible for the first time to determine statistically reliable voice parameters of adolescent speakers for German. This has been demonstrated by analyzing the f0 variability of 762 speakers and different utterance types. Linking f0 variability to utterance types provides insights into human organization of speech in real-world applications, and has implications for speech recognition technology and spoken dialog system design – by restricting the user input to utterances with less variability, the robustness of dialog systems may be increased.

Acknowledgements

The authors would like to thank the pupils and schools for their participation in the recordings, the BITS project team for the collection of the Ph@ttSessionz speech database, and the many students for their transcription efforts. The project was funded by BMBF under grant no. 01IVB01.

6. References

S. Bennett. 1983. A 3-year longitudinal study of school-aged children's fundamental frequencies. *Journal of Speech and Hearing Research*, 26:137–142, March 1983.

P. Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

S. Burger and F. Schiel. 1998. RVG 1 – A Database for Regional Variants of Contemporary German. In *Proc. LREC*, pages 1083–1087, Granada.

Chr. Draxler and K. Jänsch. 2004. Speechrecorder – a universal platform independent multi-channel audiorecording software. In *Proc. of LREC*, pages 559–562, Lisbon.

Chr. Draxler and K. Jänsch. 2006. Speech recordings in public schools in Germany - the perfect show case for web-based recordings and annotation. In *Proc. of LREC*, Genova.

U. Hollmach. 2003. Untersuchungen zur Kodifizierung der Standardaussprache in Deutschland. Technical report, Universität Halle.

J. Huber, E. Stathopoulos, G. Curione, T. Ash, and K. Jonson. 1999. Formants of children, women, and men: The effects of vocal intensity variation. *Journal of the Acoustical Society of America*, 106(3):1532–1542, Sept.

T. Hudson, G. de Jong, K. McDougall, P. Harrison, and F. Nolan. 2007. F0 Statistics for 100 Young Male Speakers of Standard Southern British English. In *Proc. of ICPhS*, pages 1809–1812, Saarbrücken.

P. Keating and R. Buhr. 1978. Fundamental frequency in the speech of infants and children. *Journal of the Acoustical Society of America*, (62):567–571, February.

R. Kent. 1976. Anatomical and neuromuscular maturation of the speech mechanism: evidence from acoustic studies. *Journal of Speech, Language, Hearing Research*, 19(3):421–447.

G. Kovacic and A. Budanovac. 2002. Acoustic characteristics of adolescent actor's and non-actors' voices. *Folia Phoniatrica et Logopaedica*, (54):125–132.

H. Künzel. 1989. How well does average fundamental frequency correlate with speaker height and weight? *Phonetica*, 46:117–125.

S. Lee, A. Potamianos, and S. Narayanan. 1999. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *Journal of the Acoustical Society of America*, 105(3):1455–1468, March.

J. Lindh. 2006. Preliminary descriptive f0-statistics for young male speakers. Technical report, Working papers, Dept. of Linguistics & Phonetics, Centre for Languages & Literature, Lund University, Lund.

F. Nolan, K. McDougall, G. de Jong, and T. Hudson. 2006. A forensic phonetic study of 'dynamic' sources of variability in speech: The dyvis project. In *Proc. of SST*, Auckland.

F. Senia and J. van Velden. 1997. Specification of Orthographic Transcription and Lexicon Conventions. Technical Report SD1.3.2, SpeechDat-II LE-4001.

S. Whiteside and C. Hodgson. 2000. Some acoustic characteristics in the voices of 6- to 10-year-old children and adults: a comparative sex and developmental perspective. *Logopedics Phoniatrics Vocology*, 25:122–132.

S. Whiteside, C. Hodgson, and C. Tapster. 2002. Vocal characteristics in pre-adolescent and adolescent children: a longitudinal study. *Logopedics Phoniatrics Vocology*, 27:12–20.

R. Winski. 1997. Definition of Corpus, Scripts, and Standards for Fixed Networks. Technical report, SpeechDat Report LE2-4001-SD1.1.1.