A Database of Age and Gender Annotated Telephone Speech

Felix Burkhardt, Martin Eckert, Wiebke Johannsen, Joachim Stegmann
Deutsche Telekom AG Laboratories
Ernst-Reuter-Platz 7, 10587 Berlin, Germany
E-mail: Felix.Burkhardt | Martin.Eckert | Wiebke.Johannsen | Joachim.Stegmann@telekom.de

Abstract

This article describes an age-annotated database of German telephone speech.
We recorded 47 hours of prompted and free text, spoken by 954 paid participants in a style typical for automated voice services.
The participants were selected based on an equal distribution of males and females within four age cluster groups.
The data will be used to tune an age-detecting automated voice service and might be released to research institutes under controlled conditions as part of an open age and gender detection challenge.

1. Introduction

Speech based classification has been gaining more and more attraction in the past years.

In contrast to speech recognition, the term refers to the detection of speaker of short or longer term characterizations like age, gender, language, sociolect, dialect, accent, health, emotion, mood or personality.

In human-human communication speaker classification takes place at all times and is very valuable for the communication process, as people constantly adapt their manner of speaking based on the assessment they have of their counterpart’s age, gender, mood or mother tongue.

Incorporating such strategies into (semi-)automated voice services can be very helpful to extend their benefit.

In (Burkhardt et al, 2007) we motivate the use of age and gender detection in several applications regarding human machine dialog interaction.

Some applications shall be listed here to illustrate ideas:

- Adapting dialog/persona design:
- Target group specific advertising
- Market research
- Gaming/Fun

A first pilot study introducing speech based classification concerns customer research with respect to tuning data of a customer self service portal of the German Telekom.

The age structure of the calling customers should be detected automatically by speech analysis.

Because all existing technical approaches to detect age from voice are based on statistical methods, a data collection of age and gender annotated telephone speech is vital.

We started using existing databases for German telephone speech like the German Speech Dat corpus (Höge et al, 1999) and described outcomes for automatic age and gender classification in (Metze et al, 2007).

The German Speech Dat data has the advantage to include a large number of speakers (4000), but firstly it is not focused on age detection and therefore the age structure is not very balanced, and secondly the textual content diverges from the utterances likely to appear in the target voice portal.

For this reasons we undertook an own data collection focussed on age and gender which gets described in this article.

The data collection should be

- on the one hand similar enough to the Speech Dat corpus to be used as an extension and gain comparable results on the two corpora,
- on the other hand near enough to the data stemming from the target application, i.e. the German Telekom service portal, as to be used as a training corpus for the application.

In the following section we describe firstly the methodology we used to record the data.

A next section is about the criteria to select the source speakers, followed by a section on the textual content of the data.

We describe then the resulting database, followed by a short section on additional data stemming from earlier projects.

A last section reflects on the article and mentions next steps.

2. Methodology

An external company was assigned to identify possible speakers of the targeted age- and gender groups (see section 3).

The participants received written instructions on the procedure and got a financial reward for their participation, the calls were free of charge.

They were asked to ring up the recording system six times.

Each time they were prompted by an automated Interactive Voice Response (IVR) system to repeat given utterances or speak free content (section 4).

The speakers got individual prompt sheets containing the utterances and additional instructions.

Between each session a break of one day was scheduled to get more variations of the voices.

Since the dropout rates at such recording sessions are quite high, i.e. quite often the participants stop the recording session before all questions are answered, the recording system did not check the breaks to avoid losing too much speakers.

Six calls had to be done with a mobile phone alternating indoor and outdoor to obtain different recording environments.

The caller is connected by mobile network or ISDN and PBX to the recording system.

The recording system consists of an application server hosting the recording application and a VoiceXML telephony server (Genesys Voice Platform, GVP).

During the call the utterances were recorded by the VoiceXML interpreter using the recording feature provided by the voice platform and sent to the application server.

The utterances were stored on the application server as 8 bit, 8 kHz, A-law.

To validate the data, the associated age cluster (see section 3) was compared with a manual transcription of the self stated date of birth in question 10 (see section 4).

3. Speaker selection

Four age groups, including children, young-aged, middle-aged, and seniors, were defined for the speech database.

The choice was not motivated by any physiological aspects that might arise from the development of the human voice with increasing age, but solely on market aspects stemming from the application.

Since children are not subdivided into male and female, this results in seven classes as shown in table 1.

class	**age group**	**age**	**gender**
#1	Children	7 – 14	m + f
#2	Young	15 - 24	f
#3	Young	15 - 24	m
#4	Middle	25 - 54	f
#5	Middle	25 - 54	m
#6	Seniors	55 - 80	f
#7	Seniors	55 - 80	m

Table 1: Age and gender classes

The following paragraphs describe the requirements that were communicated to the company assigned with the speaker recruitment.

It was planned to record at least 100 German speakers for each class.

These speakers should be acquired from all German Federal States but it was not necessary to perfectly balance out the distribution of German dialects.

The participation of multiple speakers from one household was allowed.

For the children class an age of seven years was given as lower limit.

The ability to read the given phrases was a precondition for the participation of children.

The upper age limit for the seniors class was 80.

Within one class the speakers age was intended to be uniformly distributed.

That means, ideally, that for each age (or year of birth) within a class the same number of speakers should be contained.

As a minimum requirement in this context we defined age sub-clusters of equal size.

To account for the different age intervals of the groups, children and young-aged should be uniformly distributed within two year clusters and middle-aged and seniors in five year clusters.

For example this means that 25 children from seven to eight years and 20 young-aged females between 17 or 18 years should participate.

All age groups, including the children, should have equal gender distribution.

4. Text selection

The content of the database was designed in the style of the Speech Dat corpora.

The data collection should be on the one hand similar enough to the Speech Dat corpus to be used as an extension and gain comparable results on the two corpora, on the other hand near enough to the data stemming from the target application, i.e. the German Telekom service portal, as to be used as a training corpus for the application.

Each of the six recording sessions contained 18 utterances taken from a set of utterances listed in the table below.

#	**topic**	**quantity / description**	**examples / motivating questions**
1-3	**command word**	preset words 18 different command words	“Stop” “Hilfe” (help)
4-5	**embedded command**	preset phrases 22 different command words embedded in 2 to 5 phrases each	“einen *Berater* bitte“ (an *agent* please)
6	**month**	preset words all 12 months	“Januar” (January) “Dezember”

			(December)
7	**week day**	preset words all 7 days	“Montag” (Monday) “Sonntag” (Sunday)
8	**relative time description**	preset words 9 different words/phrases	“heute” (today) “nächste Woche” (next week)
9	**public holiday**	preset words 6 different words	“Ostern” (Easter) “Weihnachten” (Christmas)
10	**birth date**	free wording	“When is your birthday?”
11	**time**	free wording	“What time is it?”
12	**date**	free wording	“Please tell us any date, for example the birthday of a family member.”
13	**telephone number**	Free wording	“Please tell us – number by number – any phone number, for example your own call number.”
14	**postcode**	Free wording	“Please tell us your post code.”
15	**first name**	Free wording	“Please tell us your first name.”
16	**last name**	Free wording	“Please tell us your last name.”
17-18	**yes / no**	Free wording	“Is it raining just now?” “Are you outside at the moment?”

Table 2: Content of speech database

Each participant had to call the recording system six times.

On an accompanied instruction sheet all items, relevant for the specific recording session, where listed.

To motivate the free wording content the above mentioned questions were asked.

There was no control that the personal dates, like names or birthdates, correspond with the information given at the screening of the callers.

Furthermore there was no need to repeat the same names on every recording session.

Within the set of the preset words it was attended that the content for each speaker did not recur.

5. Results

All in all we collected data of 954 speakers. In figure 1, , the histogram of the age structure is shown.

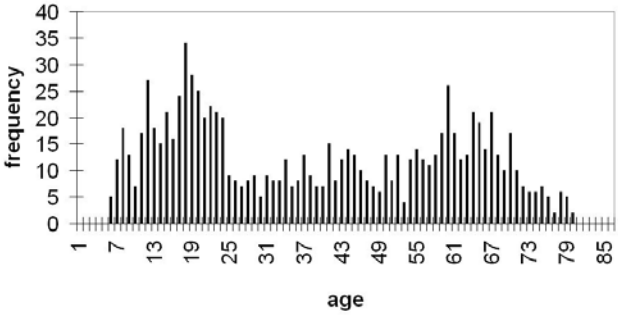

Figure 1: Age structure

Not all speakers finalized all six calls, while some even called more often than six times, so there are different numbers of utterances for each speaker.

All in all we collected 65364 single utterances.

They were in average of 2.58 seconds duration, all in all the database comprises 47.008 hours of speech.

Of those, we selected for each of the seven classes randomly 25 speakers as a fixed test set (17332 utterances, 12.45 hours) and the other 770 speakers as a training set (53076 utterances, 38.16 hours).

The following table lists the number of speakers and the number of utterances per class in the training set.

Class	**1**	**2**	**3**	**4**	**5**	**6**	**7**
#spk.	106	99	88	113	107	123	134
#utt.	6804	7360	6189	7844	6911	8575	9575

Table 3: Number of speakers per class in training set

6. Additional databases

From an earlier project, the “VoiceClass” database was collected.

It consists of 659 speakers (368 female and 291 male) living in Berlin that called an automated voice portal server and answered freely on of two questions:

- “What is your favourite dish?” (food set., 237 speakers)
- “What would you take to an island?” (island set, 422 speakers)

The age range of the data is from age four to seventy five, non-uniformly distributed, the duration of the utterances ranges from 2.4 to 98.3 seconds.

The age structure shown in figure 2 reveals that children are over-represented in the data.

We used this database for out-of domain testing, i.e. in order to benchmark classifiers that were trained on other data, like described in (Metze et al, 2007).

In real applications, we quite often have the situation that no initial training data is available and we have to rely on out of domain training data.

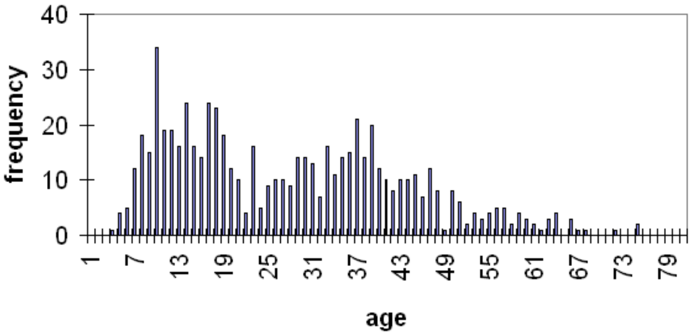

Figure 2: Age structure of the VoiceClass database

7. Conclusions and outlook

This article described an age-annotated database of German telephone speech.

We recorded 47 hours of prompted and free text by 954 paid participants in a style typical for automated voice services.

A second database might be used for out of domain age recognition evaluation.

The databases has been created for three purposes.

On the one hand we use it to train the existent pilot application that estimates customer’s age.

Secondly features and classifiers can be evaluated based on the data. A first evaluation is described in (Müller et al, 2009).

Thirdly, the release of the data to research institutes under controlled conditions as part of an open age and gender detection challenge database is planned.

8. References:

Burkhardt, F., Huber, R. and Batliner, A. (2007): Application of Speaker Classification in Human Machine Dialog Systems, in Speaker Classification I: Fundamentals, Features, and Methods, edited by C. Müller, Springer, pp 174-179

Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., Müller, C., Huber, R., Andrassy, B., Bauer, J.G., Little, B. (2007): Comparison of Four Approaches to Age and Gender Recognition for Telephone Applications, Proc. ICASSP

Müller, C., Feld, M. and Burkhardt, F. (2010): GMM-SVM Supervector Approach on Speaker Age Recognition for Short Telephone-Quality Utterances, to appear

Höge, H., Draxler, C., van den Heuvel, H., Johansen, F. T., Sanders, E., Tropf, H. S. (1999): SpeechDat Multilingual Speech Databases for Teleservices: Across the Finish Line, Proc EUROSPEECH