The Influence of the Utterance Length on the Recognition of Aged Voices

Alexander Schmitt[1], Tim Polzehl[2], Wolfgang Minker[1], Jackson Liscombe[3]

[1]Institute for Information Technology, University of Ulm, Germany
[2]Quality and Usability Lab TU Berlin / Deutsche Telekom Laboratories, 10587 Berlin, Germany
[3]SpeechCycle Inc., Broadway 26, New York, USA
alexander.schmitt@uni-ulm.de, tim.polzehl@telekom.de, wolfgang.minker@uni-ulm.de, jackson@speechcycle.com

Abstract

This paper addresses the recognition of elderly callers based on short and narrow-band utterances, which are typical for Interactive Voice Response (IVR) systems. Our study is based on 2308 short utterances from a deployed IVR application. We show that features such as speaking rate, jitter and shimmer that are considered as most meaningful ones for determining elderly users underperform when used in the IVR context while pitch and intensity features seem to gain importance. We further demonstrate the influence of the utterance length on the classifier's performance: for both humans and classifier, the distinction between aged and non-aged voices becomes increasingly difficult the shorter the utterances get. Our setup based on a Support Vector Machine (SVM) with linear kernel reaches a comparably poor performance of 58% accuracy, which can be attributed to an average utterance length of only 1.6 seconds. The automatic distinction between aged and non-aged utterances drops to random when the utterance length falls below 1.2 seconds.

1. Introduction

Spoken dialogue technology has reached such a sophisticated level that it enables a growing complexity of telephone-based speech applications. Richer speech input possibilities allow for another generation of the so-called Interactive Voice Response (IVR) applications: the automated troubleshooters (Acomb et al., 2007). Unlike previous generations of IVR systems being of information retrieval or transactional character, these systems provide automated technical support and guide callers jointly towards a solution of their problem. By that, thousands of Internet- or television-related problems are solved every day through automation.

The complexity of those systems has risen substantially. While early IVR systems consisted of only few dialogue steps, problem solving applications may contain several dozen and frequently up to 50-100 dialogue steps in one call. Both, customers and providers have a substantial interest that the call in which they invested a substantial amount of time and money ends up successfully. While the customer is at risk of uselessly spending time with a system that might not solve her problem, the provider in return has running costs for each call that blocks a port on his rented telephone platform. Even more severe may be the loss of corporate image for the carrier when the system dissatisfies the caller.

Task completion is the highest precept in this structure and every situation has to be avoided where callers prematurely and unsatisfied put down the receiver.

2. Motivation

In earlier work we have presented a problematic dialogue predictor (PDP), i.e. a statistical classifier, being able to detect problematic dialogues based on interaction logs (Schmitt et al., 2008). The target application is that of an automated Internet troubleshooter. The classifier's prediction is entirely based on log data such as the number of Automatic Speech Recognition (ASR) errors, average number of times the user did not respond in time, average barge-in rate etc. An extensive set of linguistic and discourse information available from the log data is included in the classification. An anger detection system (Schmitt et al., 2009a) is envisioned to be incorporated and will potentially raise the PDP's accuracy. However, further details that can be exploited from the caller's voice such as age and gender have been neglected so far.

Analyzing the task completion rate of the automated Internet troubleshooter, an IVR application helping callers to resolve Internet-related problems, we found out that non-senior callers had a 33% higher task completion rate than senior callers. An adaption of the dialogue to elderly callers, could potentially rise their task completion rate, e.g. by introducing more explicit confirmations or in case of additional problems detected by the PDP, an immediate transfer to an operator who can help out.

Other application scenarios for a distinction between senior and non-senior callers are:

- A shifting of the acoustic ASR models to models that are especially trained on aged voices to raise the recognizer's accuracy.
- Self-service applications that employ advertisements tailored to the specific user group while the caller is on hold for a live operator.

3. Related Work

Humans are able to estimate the age of speakers in rough dimensions (Minematsu et al., 2002). Certainly, distinguishing between e.g. males being aged 35 or aged 45 is virtually impossible, but a reasonable distinction between children, young adults and seniors is feasible.

However, *automated* speech-based age recognition, presumably because of its difficulty, is still in an early phase. There exist only few studies that deal with the recognition of speaker age and fewer that consider short narrow-band utterances which are typical for IVR systems. (Müller et al., 2003) present a study on the recognition of aged voices. They employ 5 jitter and 3 shimmer features to determine elderly from non-elderly voices by using a Bayesian Network. The corpus employed in the tests has been merged from two different corpora which could potentially have led to a bias in the recognition results. The presented values are difficult to judge due to the fact that the corpus is strongly unbalanced showing a higher degree of younger speakers than older ones.

(Bocklet et al., 2008) present a study on classifying children, young adult, adult and senior speakers based on two different telephone corpora. They employ Gaussian Mixture Models and Support Vector Machines for classification. On the first corpus (SpeechDat II), f1 scores of about 75.5% could be reached, on the second corpus (VoiceClass) approximately 60.5%. While this can be considered as a satisfying result for a five-class recognition problem, it is important to note that the speech samples are of considerable length (5-30 seconds in VoiceClass) and are not comparable to typical IVR utterances.

Unlike the cited studies, we analyze the performance of different acoustic and prosodic feature groups on the classification result. Additionally we direct our attention to the influence of the utterance length on the age recognition task.

4. Corpus Design

For our study we analyzed 1,911 calls from the automated Internet troubleshooter. The disadvantage of "real-life" data is that the speaker age is mostly unknown, since it is generally difficult or rather impossible to ask customers to enter their age when using IVR systems. Instead, a manual rating is required. We asked three expert raters to label each call according to the labels "younger than 60", "older than 60" and "unsure". Consequently, the perceptual age is considered rather than the real age and "60" can only be considered as a rough "division point".

Following the assumption that it is easier for raters to judge over the speaker age when listening to longer utterances rather than to shorter ones, we sorted each call internally according to the utterance length. We started the labeling process by presenting to the raters the call containing the longest *first* utterance in the corpus and ending with the call containing the shortest *first* utterance. The procedure is depicted in Figure 1. Thereby the raters listened first to the longest utterance of each call and the given label was assigned to all other utterances in the call which sped up rating considerably. When the rater was unsure, she could request the second longest utterance, the third longest etc.

Garbage turns (labeled in a previous labeling session) were sorted out. All three raters had a fair agreement ($\kappa = 0.21$). The final label was assigned according to majority voting. In 95 calls, no agreement between all three raters was achieved and the calls along with 1,005 utterances were sorted out. The resulting distribution was 18,550 utterances from non-senior callers, 1,664 from senior-callers (1,034 male, and 627 female), 380 unknown and 1,167 rubbish utterances.

In order to prevent data skewness and by that a bias towards a distinct age class we created a balanced subset for training and testing purposes. Furthermore we evaluated on a gender-dependent level which allows to exclude gender-dependent differences. The resulting classes are young male (YM), young female (YF), senior female (SF), senior male (SM) each consisting of 577 utterances from 55 speakers.

5. Classification

5.1. Acoustic and Prosodic Features

The most promising acoustic and prosodic features discussed in literature to determine speaker age of adults are jitter, i.e. the perturbations in pitch, and shimmer, i.e. the perturbation in power. Both account for the fact that during aging the muscle of the vocal fold loses bulk and the flexible tissues which are responsible for vocal fold vibration during voicing become stiffer and less elastic. Using the acoustic software PRAAT (Boersma and Weenink, 2009), we have extracted from each utterance 5 jitter values (*jitter_local, jitter_local_absolute, jitter_rap, jitter_ppq5, jitter_ddp*) and 5 shimmer values (*shimmer_local, shimmer_local_db, shimmer_apq3, shimmer_apq5, shimmer_apq11, shimmer_dda*) as described in (Müller et al., 2003). Under the assumption that elderly speakers tend to speak more slowly than younger speakers, the speech rate is frequently considered as important feature:

$$\frac{s}{d}$$

or

$$\frac{s}{d-p}$$

, where s is the number of syllables in an utterance, d the total duration of the utterance and p the pauses between the speech parts. We extracted both speech rates with the aid of the syllable detection script provided for PRAAT (De Jong and Wempe, 2009). Finally, pitch can be considered as a very important feature since elderly voices appear often sharply.

Moreover we have added a variety of other acoustic and prosodic features to our training set, which we have designed previously for an anger detection task (Schmitt et al., 2009b). For the anger detection system we calculated for each utterance Mel Frequency Cepstral Coefficients (MFCCs), pitch, harmonicity, formants, intensity and power. Each of these feature groups consists of the core values and their means, first and second order derivation, extrema and ranges. All features discussed here have been extracted from the complete utterance, i.e. no framing has been applied beforehand. The overall number of features per utterance used for analysis amounts to 72.

5.2. Experimental Setup

Since computational issues have to be considered when designing an age recognizer, we have analyzed the impact of the respective feature groups on the classifier's result. As classifier we have applied a Support Vector Machine (SVM) with linear kernel which was trained on the features of each feature group. In order to exploit the maximum available data, we applied leave-one-speaker-out (LOSO) classification. Hereby we performed n iterations, while n being the number of speakers, trained the SVM with all utterances from $n-1$ speakers and tested with the utterances of the remaining speaker. The obtained predictions along

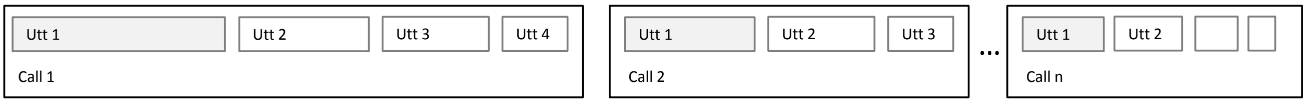


Figure 1: Sorting the corpus prior to rating: the utterances within a call are ordered according to their length in descending order. The call with the longest *first* utterance is presented first, the one with the shortest *first* utterance is presented last.

with the corresponding classes were used to determine the overall accuracy.

5.3. Analysis of Feature Groups

Figure 3 depicts the average accuracy values for the respective feature groups. Please note that we have created a completely balanced test set and therefore accuracy constitutes a reliable evaluation criterion. Each accuracy score denotes the performance of the SVM when trained and tested *only* with the respective features of a group evaluated with LOSO. Obviously, gender-related differences can be considered as negligible. The single features perform similarly in all three scenarios. It does not play a role whether the model was trained and evaluated on male, female or mixed samples, which suggests a gender-independent age recognizer. Merely intensity shows slight differences between male and female speakers.

Interestingly, the most promising feature groups for detecting aged voices, jitter, shimmer and speech rate perform very poor. The speech rate without pause performs worse than the speech rate that has been calculated on the complete speech- and non-speech parts of the utterance. Comparably best, but still not satisfying, is the performance of the intensity feature group. Pitch and intensity estimations seem to work better on short utterances. The speech rate presumably only unfolds its expressiveness when utterances with more words are to be considered.

The fact that all utterances incorporated with the corpus are short utterances potentially explains this result. Most utterances consist of one word (e.g. “yes”, “no”, “continue”), while only few contain several words (cf. Section 6.). A histogram of the utterance durations is depicted in Figure 2.

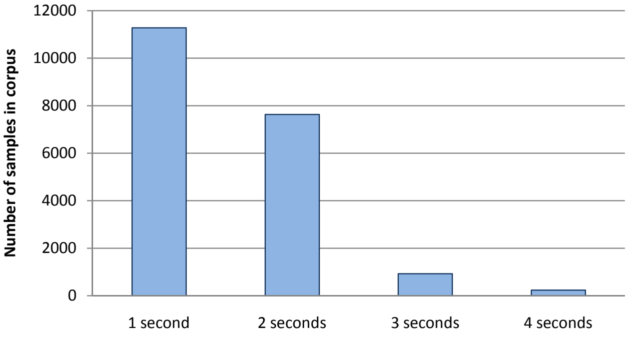


Figure 2: Distribution of utterances with respect to utterance duration (rounded). Utterances shorter than 0.5 seconds and longer than 4.5 seconds are not displayed since their occurrence is neglectible.

6. Influence of Utterance Length on Specific Feature Groups

Estimating speaker age becomes increasingly difficult the shorter the utterances are. We could observe this phenomenon during the rating process. As described in Section 4. the raters were presented the calls in descending order, i.e. the calls in the beginning of the rating process contained longer utterances than at the end of the process. The growing uncertainty can be observed in Figure 4. Subfigure 1 presents the number of calls labelled as non-seniors.

It is interesting to note that regardless of the utterance length the raters constantly labeled calls as non-seniors. Obviously the decision on rating utterances as “non-senior” was not affected by the length. Rating senior speakers seems to be a less trivial task. With decreasing utterance length, the raters labeled less calls as ’senior’ and chose ’unsure’ instead.

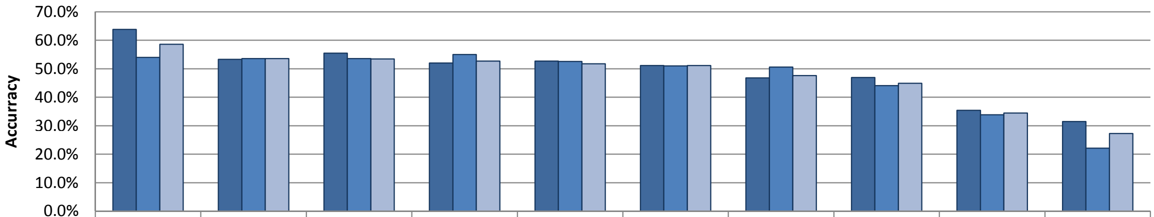
This and the results from Section 5.3. motivated us to analyze the influence of the utterance length on the automatic classifier. Under the assumption that the utterance length affects the performance of only some acoustic and prosodic features in our data, we considered the correctly classified utterances in our LOSO experiment with respect to each feature group and the utterance length.

One would expect that features that can reliably be determined on short utterances such as pitch and intensity perform equally well on both short utterances as well as on longer utterances. Furthermore, features that should be more meaningful when calculated from longer utterances, such as the speaking rate should show an increasing performance on longer utterances. We analyzed the classifier’s prediction from our feature subgroup experiment from Section 5.3. according to the utterance length. We divided the corpus of 2308 samples into subsets that contain utterances that lastet 0-1s, 1-2s, 2-3s, 3-4s, 4-5s. For each group we counted the number of correctly classified and incorrectly classified samples obtaining an accuracy of the classifier with respect to the utterance length and each specific feature group. Results are depicted in Figure 5.

Against our hypothesis, the performance of pitch and intensity show a dependency with respect to the utterance length. The performance of jitter, harmonicity and the formants is not affected by the utterance length in contrast to power, shimmer, the speech rate, intensity, pitch, tmax and the MFCCs. The latter feature groups gain substantial performance when the duration of the turns exceeds 4 seconds.

7. Overall performance

To provide a baseline for comparisons we performed classification on *all available features* with speaker-independent 10-fold cross validation on the SVM. The classifier



	intensity	MFCCs	formants	pitch	tmax	speechrate	jitter	speechrate w/o pause	harmonicity	shimmer
■ male	63.8%	53.3%	55.5%	52.0%	52.7%	51.1%	46.8%	47.0%	35.4%	31.5%
■ female	54.0%	53.6%	53.6%	55.0%	52.6%	51.0%	50.6%	44.1%	33.8%	22.1%
□ mixed	58.6%	53.6%	53.5%	52.7%	51.8%	51.2%	47.6%	44.9%	34.4%	27.3%

Figure 3: Average accuracy values of LOSO classification for respective *feature groups* on a SVM with linear kernel

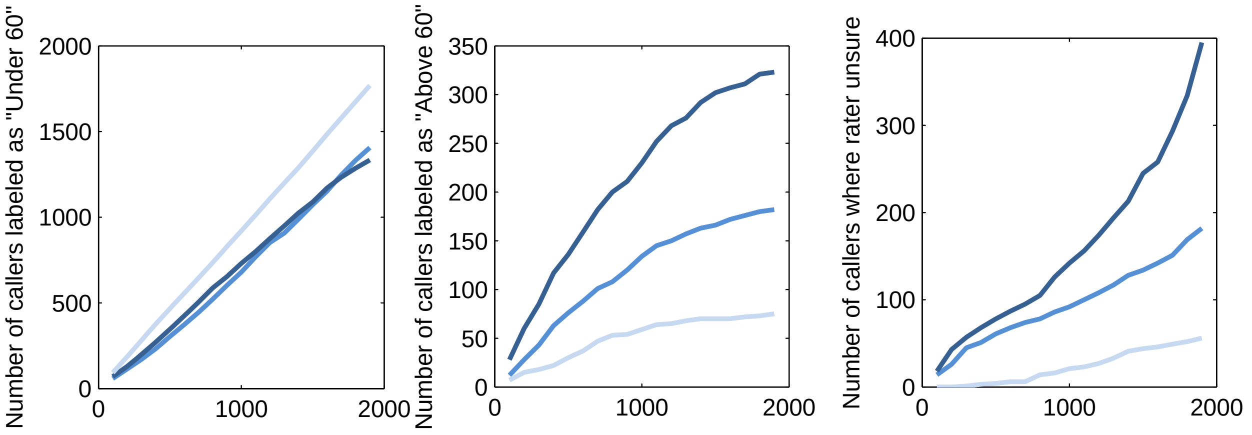


Figure 4: 3 Raters annotating age of 1,911 callers choosing between ">60", "<=60" and "unsure". First plot: number of calls the rater annotated with '<=60'. Second plot: number of calls each rater annotated with '>60', third plot: number of calls where the raters were unsure. Note that the length of the presented sample decreases. As can be seen, with decreasing utterance length the uncertainty increases and the raters favor for 'unsure' instead of '>60'.

achieved a performance of 58.09% +/- 6.54%, which appears rather weak. Considering the fact that the average duration of the classified utterances amounts to 1.6s, however, this can be still considered as acceptable. The approach of operating on very short utterances when classifying speaker age has to be questioned in general, instead the certainty would substantially rise when concatenating several turns.

8. Conclusion and Discussion

In this study we analyzed the performance of acoustic and prosodic feature groups when applied in a senior-/non-senior recognition task using short utterances of narrowband quality from an Interactive Voice Response system. The study has been carried out on a completely balanced and comparably large dataset and was evaluated speaker-independent with LOSO evaluation.
The findings are:

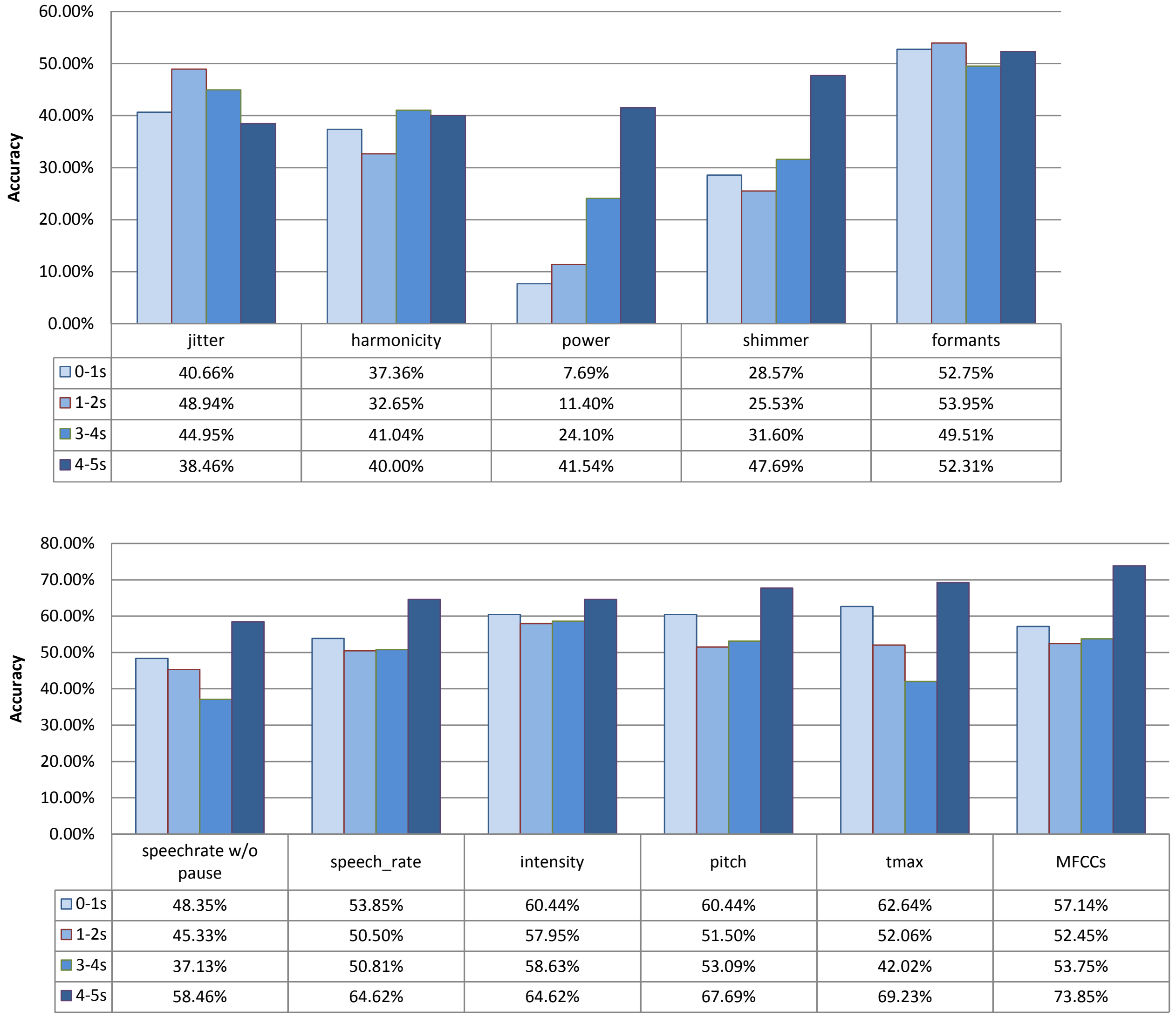
- The employed discriminative classifier, a SVM with linear kernel performs best when trained on pitch and intensity and not, as expected on jitter and shimmer.

- The speech rate fails on short utterances. The speech rate determined when including pauses outperforms the speech rate determined without pauses.

- Human annotators have big difficulties in estimating the speaker age with decreasing utterance length, particularly when rating senior voices.

- The classifier had similar problems as the annotators: a strong influence of the utterance length on the overall performance could be observed. The classifier's performance dropped to random when the utterance length was shorter than 1.2s.

An isolated view on the utterance duration with respect to the single feature groups shows that some groups improve with increasing length while others are not affected by the duration. A critical point especially in the analysis of duration in conjunction with specific feature groups is the lack of sufficient speech samples of longer lasting utterances.
According to our findings an early prediction of the speaker age in a deployed IVR system may not deliver robust results. The prediction would be too uncertain due to a lack of speech samples from the caller. The results suggest a concatenation of several short user utterances that have been captured so far during system usage prior to classification. This will increase robustness and a higher accuracy in the classification process. An online decision on the speaker age should thus be postponed until enough data is gathered.



	jitter	harmonicity	power	shimmer	formants
0-1s	40.66%	37.36%	7.69%	28.57%	52.75%
1-2s	48.94%	32.65%	11.40%	25.53%	53.95%
3-4s	44.95%	41.04%	24.10%	31.60%	49.51%
4-5s	38.46%	40.00%	41.54%	47.69%	52.31%

	speechrate w/o pause	speech_rate	intensity	pitch	tmax	MFCCs
0-1s	48.35%	53.85%	60.44%	60.44%	62.64%	57.14%
1-2s	45.33%	50.50%	57.95%	51.50%	52.06%	52.45%
3-4s	37.13%	50.81%	58.63%	53.09%	42.02%	53.75%
4-5s	58.46%	64.62%	64.62%	67.69%	69.23%	73.85%

Figure 5: Percentage of correctly classified utterances within selected feature groups analyzed according to the utterance length. Jitter, harmonicity and formants don't gain performance with increasing utterance length. In contrast, substantial increase can be observed when the utterance lasts at least 4 seconds which affects particularly power, shimmer, intensity, pitch, tmax and MFCCs.

Given the results from Figure 5 we may carefully conclude that a performance boost can be expected with turns that last at least 4 seconds.

It has to be clarified to which extent the narrow-band quality has an influence on the performance of the single features. To foster these findings, a larger study with multiple corpora is necessary.

9. Acknowlegdements

This research has been carried out in the context of the Transregional Collaborative Research Centre SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” funded by the German Research Foundation (DFG).

10. References

Kate Acomb, Jonathan Bloom, Krishna Dayanidhi, Phillip Hunter, Peter Krogh, Esther Levin, and Roberto Pieraccini. 2007. Technical support dialog systems:issues, problems, and solutions. In *Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, pages 25–31, Rochester, NY, April. Association for Computational Linguistics.

Tobias Bocklet, Andreas Maier, Josef Bauer, Felix Burkhardt, and Elmar Nöth. 2008. Age and gender recognition for telephone applications based on gmm supervectors and support vector machines. In IEEE Computer Society Press, editor, *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 1605–1608.

Paul Boersma and David Weenink. 2009. Praat: doing phonetics by computer (version 5.1.04), April.

N. H. De Jong and T. Wempe. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2):385–390, May.

Nobuaki Minematsu, Mariko Sekiguchi, and Keikichi Hirose. 2002. Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers. In *Proc. of ICASSP*, volume 1, pages 137–140.

Christian Müller, Frank Wittig, and Jörg Baus. 2003. Exploiting speech for recognizing elderly users to respond to their special needs. In *Proceedings of the Eighth European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 1305 – 1308.

Alexander Schmitt, Carolin Hank, and Jackson Liscombe. 2008. Detecting Problematic Calls With Automated Agents. In *4th IEEE Tutorial and Research Workshop Perception and Interactive Technologies for Speech-Based Systems*, Irsee (Germany), June.

Alexander Schmitt, Tobias Heinroth, and Gregor Bertrand. 2009a. Towards emotion, age- and gender-aware voicexml applications. In *5th International Conference on Intelligent Environments (IE09)*, July.

Alexander Schmitt, Tobias Heinroth, and Jackson Liscombe. 2009b. On nomatchs, noinputs and bargeins: Do non-acoustic features support anger detection? In *Proceedings of the SIGDIAL 2009 Conference*, pages 128–131, London, UK. Association for Computational Linguistics.