Speech Input from Older Users in Smart Environments: Challenges and Perspectives

Ravichander Vipperla, Maria Wolters, Kallirroi Georgila, and Steve Renals

[1] Centre for Speech Technology Research, School of Informatics, University of Edinburgh
[2] Institute for Creative Technologies, University of Southern California
r.c.vipperla@sms.ed.ac.uk, maria.wolters@ed.ac.uk,
kgeorgila@ict.usc.edu, s.renals@ed.ac.uk

Abstract. Although older people are an important user group for smart environments, there has been relatively little work on adapting natural language interfaces to their requirements. In this paper, we focus on a particularly thorny problem: processing speech input from older users. Our experiments on the MATCH corpus show clearly that we need age-specific adaptation in order to recognize older users' speech reliably. Language models need to cover typical interaction patterns of older people, and acoustic models need to accommodate older voices. Further research is needed into intelligent adaptation techniques that will allow existing large, robust systems to be adapted with relatively small amounts of in-domain, age appropriate data. In addition, older users need to be supported with adequate strategies for handling speech recognition errors.

1 Introduction

Older people are an important user group for many types of smart environments, ranging from sophisticated home automation systems to state-of-the-art environmental control systems. Speech can form an important interface for smart home environments because it is hands-free and enables potentially richer interactions. Spoken interaction is of particular benefit for people with mobility restrictions, such as those caused by diseases such as rheumatism and arthritis which affect one in three adults over the age of 65 in the UK[1]. Speech input and output is also very useful for visually impaired people: 10% of the population aged 65-74 in the UK is visually impaired[2].

Although there has been an increasing amount of research in smart home environments [1], there has been limited use of speech-based interactions. This is largely due to the challenges posed by spoken language systems in domestic environments. If the users are not forced to wear microphones, or to interact via some kind of handset, then room-based microphones distant from the user must be used. This dramatically

[1] National Statistics: Morbidity: Arthritis is more common in women. http://www. statistics. gov.uk/cci/nugget.asp?id=1331. Last visited 27/01/09.

[2] Rosemary Tate, Liam Smeeth, Jennifer Evans, Astrid Fletcher, Chris Owen, Alicja Rudnicka: The prevalence of visual impairment in the UK. A review of the literature. Royal National Institute for the Blind. Last retrieved 15/02/2009. www.rnib.org.uk/xpedio/groups/public/documents/publicwebsite/public_prevalencereport.doc.

C. Stephanidis (Ed.): Universal Access in HCI, Part II, HCII 2009, LNCS 5615, pp. 117–126, 2009.

increases the problem of Automatic Speech Recognition (ASR), since the users' speech must be separated from the many other acoustic sources in a home setting. Microphone arrays, which enable software directed beam forming, are an attractive approach to this problem [2], but the technology is still relatively immature, and is computationally demanding. However, good results have been achieved for large scale automatic speech recognition tasks in less demanding environments, such as business meetings [3]. Furthermore, accurate speech recognition and natural-sounding speech synthesis, do not comprise a useful interaction modality on their own. These speech technologies must be combined with speech understanding and dialogue management if a usable spoken language modality is to be provided. The INSPIRE system [4] is one of relatively few examples of a smart home system with well-developed spoken interaction.

Hands-free speech interfaces provide flexible solutions that free people from having to carry an interaction device (such as a phone) or from having to physically move to a console, and are thus well suited to older people and to people with disabilities or limited mobility. Such speech interfaces have even been shown to be feasible for users with severe speech impairments such as dysarthria, if the commands are appropriately designed and the system is sufficiently adjusted using samples of the user's speech [5]. Despite this great potential, older users' speech input presents challenges that the ASR community has only recently begun to address.

In this paper, we examine the effect of speaker age on recognition performance, in terms of acoustic variability and in terms of linguistic factors. We have performed two experiments. In the first experiment, we investigate linguistic differences between older and younger users in the context of the language modeling component of an ASR system; in the second experiment, we focus on the acoustic variability that arises from vocal ageing and report on the combined effect of linguistic and acoustic factors on ASR for older users. These experiments were performed using the MATCH corpus, which contains interactions between both older and younger users and several different appointment scheduling dialogue systems. We conclude that innovative strategies are required for adapting existing speech recognition systems to older voices; in particular to achieve high accuracy, speech recognition systems not only need to cover the precise domain of interaction, but they also have to take into account the acoustic and linguistic characteristics of older users' speech.

2 Older Speakers, Older Voices: A Challenge for ASR

The effects of ageing are notoriously difficult to study because chronological age is a relatively poor predictor of anatomical, physiological, and cognitive changes [6, 7]. This variability is not just due to genes, but also to individuals' lifestyle [8]. As a consequence, older users are notoriously difficult to design for, because individual older people will have very different needs and abilities.

With ageing, several degenerative changes occur in the respiratory system, larynx and the oral cavity which form the human speech production mechanism [9]. Significant changes affecting speech include loss of elasticity in the respiratory system leading to decreased lung pressure, calcification of the laryngeal tissues leading to the instability of the vocal fold vibrations, loss of tongue strength, tooth loss, and changes in the dimensions of the oral cavity [10]. Ageing affects many acoustic parameters of the speech wave form such as fundamental frequency, jitter, shimmer and harmonic-noise ratios [11]. Ageing voices are also characterized by increased breathiness and slower speaking rates. All these changes in the acoustics of older voices have their impact on ASR systems: Word Error Rates (WERs) for older voices are significantly higher than for younger voices [12-14].

At first blush, older users' language should not differ much from that of younger users. Even though cognitive abilities such as fluid intelligence generally decline with age [15], acquired knowledge, such as vocabulary, tends to be well-preserved [16] – to the extent that older users may use a richer vocabulary than younger ones. However, older users are more prone to word finding difficulties [17] and may produce more disfluencies under stressful conditions [18]. Word finding difficulties can lead to unexpected pauses, phrasing, and disfluencies. Patterns of word use also change during the life span. Older people use fewer words that denote negative emotions and fewer self-referential words [19].

3 The MATCH Corpus

3.1 Design and Structure of the Corpus

The MATCH corpus was recorded during a cognitive psychology experiment that investigated the accommodation of cognitive ageing in spoken dialogue interfaces [20]. 24 younger users (aged 18-29 years, mean 22) and 26 older users (aged 52-84 years, mean 66) booked health care appointments using nine different simulated spoken dialogue interfaces. Each person used each system only once in order to constrain the duration of the experiment. All dialogues were strictly system-initiative. Users could only select health care professionals and time slots proposed by the system; they were not able to suggest any aspect of the appointment themselves. This very restrictive design was chosen for two reasons: (1) it allowed us to control the dialogue structure for the purposes of the underlying cognitive psychology experiment; (2) user utterances were more likely to be restricted to the options presented in a given system message, which should make them easier to recognize.

All users participated in an extensive battery of cognitive tests before the experiment and completed detailed questionnaires rating system usability. A total of 447 dialogues were recorded using an EDIROL R01 digital recorder and a sampling frequency of 44.1 kHz.[3] The dialogues contain 3.5 hours of speech. All dialogues were transcribed orthographically by a trained annotator using the tool Transcriber [21] and the AMI transcription guidelines [22], which were used for creating the AMI meetings corpus [23]. The AMI guidelines were chosen because they have been explicitly designed to provide a solid basis for speech recognition research and to facilitate a wide range of further possible annotations. The corpus has been annotated semi-automatically with dialogue acts and information state update information [24]. For our recognition experiments, the users' speech was divided into contiguous sequences delimited by pauses, *speech spurts*. Older users produced a total of 1680 speech spurts[4] while younger users produced 1369 spurts.

[3] Recordings for three dialogues were lost.

[4] Speech spurts are contiguous sequences of user speech delimited by pauses.

3.2 Advantages and Limitations of the Corpus

Although appointment scheduling is not a central task of smart environments, it is a key functionality in many related applications such as electronic diaries and automatic scheduling of health care appointments. Since the MATCH corpus was created for a cognitive psychology experiment, dialogue structure, appointment scenarios and system vocabulary were tightly controlled. As a result, the vocabulary is much less diverse and the language is more formulaic than that of corpora which were recorded for speech research, such as DARPA Communicator [25]. It is also relatively small compared to other speech research corpora. Despite these disadvantages, the MATCH corpus is one of very few corpora that contain a large proportion of older speakers. Unlike the Dragon corpus [26] or the OYEZ corpus [27], it contains highly detailed dialogue act and information state annotations. The MATCH corpus has already been used successfully for training simulated users [28]. Simulated users typically interact with the dialogue system in order to learn dialogue policies. We found that the behavior of older users could not be modeled adequately using data from younger users – age appropriate data was needed.

3.3 Differences between Older and Younger Users

In our analyses of the MATCH data, we found substantial differences in both speech and language between older and younger users. While younger users mainly produced utterances that were directly relevant to the appointment scheduling task, older users often attempted social interaction with the system. They thanked it for providing information, or provided information that was not specified in the task definition and could not be processed by the dialogue system. In particular, older users frequently attempted to take the initiative and suggest convenient appointment slots, even though the dialogues were strictly system-initiative.

Overall, older people produce significantly more individual words (tokens) and significantly more distinct word forms (types) than younger people. Taken together the 26 older users used 373 distinct types, whereas the 24 younger users only had a vocabulary of 92 distinct types between them. Older users were more likely than younger users to use expressions other than “yes” to express agreement, such as “fine”. Older people also tend to use expressions that are more appropriate in human-human interactions, such as forms of “goodbye” or “thank you”. More detailed results can be found in [24]. These results lead us to expect that language models trained on material from younger users only will underperform when confronted with data from older users. In particular, we expect to see a high proportion of out-of-vocabulary words.

4 Experiments

In our experiments, we examined the effect of age-specific language models and acoustic models on speech recognition performance. All experiments were set up using the Hidden Markov Model Toolkit (HTK).[5]

[5] HTK version 3.4. http://htk.eng.cam.ac.uk

4.1 Experiment 1: Impact of Language Modeling

Design. The aim of this experiment was to assess the effect of the differences in interaction style between younger and older users described above, on the language modeling component of the speech recognizer and consequently on ASR performance.

From the transcripts of the MATCH corpus, the following bigram language models were constructed: 1) from all the utterances of the older speakers (*LM-Older*); 2) from all the utterances of the young speakers (*LM-Young*); 3) for each test speaker, from the entire corpus excluding the test speaker (*LM-All-1*); 4) for each older test speaker, from the corpus of all the older speakers excluding the test speaker (*LM-Older-1*); and 5) for each young test speaker, from the corpus of all the young speakers excluding the test speaker (*LM-Young-1*). For each older test speaker, three ASR experiments were performed, keeping the acoustic model fixed and using different language models for the speaker viz., LM-All-1, LM-Older-1 and LM-Young. Similarly, ASR experiments were repeated for each of the young speakers using the language models: LM-All-1, LM-Young-1 and LM-Older.

Since the amount of data in the MATCH corpus is not sufficient to build acoustic models from scratch, we used the speech from other corpora for this purpose. Acoustic models were trained on 73 hours of meetings data recorded by the International Computer Science Institute (ICSI), 13 hours of meeting corpora from the National Institute of Standards and Technology (NIST) and 10 hours of corpora from the Interactive Systems Lab (ISL) [29]. These models were then adapted using the maximum a posteriori approach [30] with 13 hours of speech from 32 UK speakers from the Augmented Multi party Interaction (AMI) data. For training the models the waveforms were parameterized into perceptual linear prediction cepstral feature vectors. Energy along with 1st and 2nd order derivates were appended giving a 39 dimensional feature vector. The acoustic models were trained as crossword context dependent triphone Hidden Markov Models (HMMs).

Results. Goodness of fit of the language model on a test set was measured using perplexity [31]. The lower the perplexity, the better is the language model for the test set. We also assessed the number of out-of-vocabulary (OOV) words, i.e. the number of words in the test set not present in the vocabulary of the language model. We found that language models trained on younger users were a bad fit of the language of older users, whereas data from the older users allowed us to model the language patterns of younger users reasonably well. In particular, models trained on younger users only did not contain many of the words older people used. These findings are consistent with the results of our experiments with simulated users discussed above. Detailed results are shown in Table 1.

Table 1. Perplexity and % OOV Words

Test Set	Language Model	Perplexity	OOV (%)
Younger	LM-Older	5.44	1.38
Older	LM-Young	19.18	15.57

Fig. 1 shows ASR Word Error Rates (WER) using different language models as explained above, averaged over all the young speakers and older speakers respectively. As we would expect from the results presented in Table 1, we find that WERs for older speakers are particularly high when using the language models of the younger speakers. This is due to the mismatch between the older and younger users' interaction styles. Clearly, we need age-appropriate data to build adequate language models for older speakers.

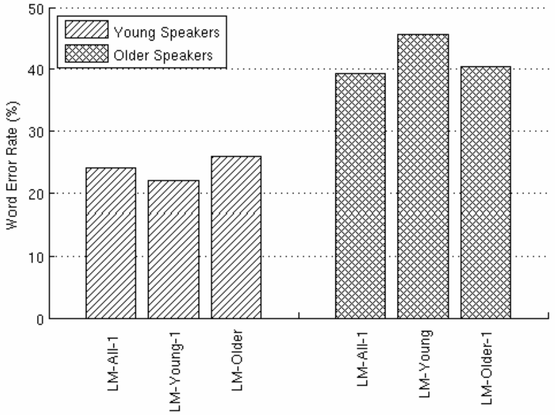

Fig. 1. ASR Word Error Rates for young and older speakers' test sets using different language models

4.2 Experiment 2: Impact of Acoustic Models

Design. In this set of experiments, we examined the impact of differences in the acoustics of older and young speakers on speech recognition performance. In order to isolate the effect of the acoustic models, we only used the language model *LM-All*, which contains all utterances in the MATCH corpus, for this set of experiments.

The acoustic models described in the previous experiment (models adapted with AMI data) were used as the baseline models. For each of the old speakers, two acoustic models were created by maximum a posteriori adaptation of the baseline models using the speech from either the rest of the old speakers excluding the test speaker (*AMI + MATCH older-1)* or speech from the young speakers (*AMI + MATCH younger*). Acoustic models were similarly created for each young speaker with the speech data from all the older speakers (*AMI + MATCH older)* and the speech data from the rest of the young speakers (*AMI + MATCH young-1).*

Results. Fig. 2 shows average WERs for both young and older speakers. The WERs for older speakers are higher than those for younger speakers by 10.99% absolute using the baseline acoustic models. Adapting the models with speech from a new domain (i.e. appointment scheduling) is expected to reduce the WERs for the test data in the new domain. While adapting the baseline models with older speakers from the

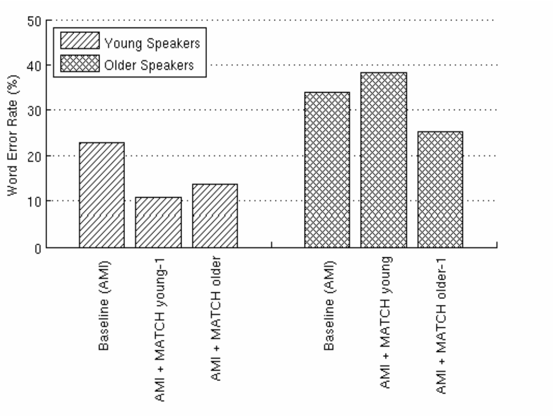

Fig. 2. ASR Word Error Rates for young and older speakers' test sets using different acoustic models

MATCH corpus (*AMI + MATCH older)* brings down the WERs for young speakers, the results are even better with adaptation using speech from other younger speakers in the same corpus (*AMI + MATCH young-1).* The results for older speakers in Fig. 2 are quite interesting, Contrary to the belief that speech from a new domain should help in creating better models for the new domain, adapting the baseline models with speech from the younger speakers of MATCH corpus (*AMI + MATCH young)* deteriorates the performance for the older speakers in the same corpus. Hence, there is a clear mismatch in the acoustics of older and young speakers resulting in a higher WER for older speakers. The reasons for this result require further investigation.

5 Conclusion

In our ASR experiments, we discovered that older users' speech resulted in higher error rates compared with the speech of younger users. This was caused by both acoustic and linguistic factors. We have performed experiments with a variety of acoustic and language models, estimated from both in-domain and out-of-domain data, derived from both younger and older users. These results have highlighted the fact that ASR systems need to take into account both acoustic and linguistic aspects of the speech of older users.

Our results indicate that the speech recognition component of a spoken dialogue system used in a smart home environment must be adapted to both the domain of usage and to the acoustic and linguistic characteristics of the users. In particular, we have shown that in-domain speech data matched to younger users does not appropriately adapt the system to the language of older users in the same domain. Even though the MATCH corpus was tightly controlled and covered a comparatively narrow domain, the findings of Möller et al. [32] suggest that we can expect to see similar results for other domains, such as controlling household items or televisions.

In order to accommodate the vocabulary and speaking patterns used by older people as well as the sound of older voices, designers and programmers need to ensure that adequate data is collected. In particular, the tasks must be clearly specified, and all relevant domains must be covered. This data set need not be large - it is possible to use existing data and small amounts of matched data to adapt generic ASR systems to task domain and user age. “Factored” adaptation algorithms are particularly promising. They can combine adaptation data that partially matches the task in question either in terms of age or in terms of domain.

Last but not least, since older people’s speech poses special challenges for ASR, systems need to provide adequate support for handling recognition errors, both within the voice modality, and in combination with other modalities.

Acknowledgements. This research was funded by SHEFC grant HR04016 MATCH (Mobilising Advanced Technologies for Care at Home). We would like to thank Robert Logie and Johanna Moore for collaborating on experiment and corpus design and analysis, Neil Mayo and Joe Eddy for coding the Wizard-of-Oz (WoZ) interface, Vasilis Karaiskos for administering the experiment, Melissa Kronenthal for transcribing all 447 dialogues, and Matt Watson for scheduling participants, administering the cognitive test battery, and data entry. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/). The ECDF is partially supported by the eDIKT initiative (http://www.edikt.org.uk).

References

1. Helal, S., Mann, W., El-Zabadani, H., King, J., Kaddoura, Y., Jansen, E.: The Gator Tech Smart House: a programmable pervasive space. Computer 38, 50–60 (2005)
2. Vovos, A., Kladis, B., Fakotakis, N.: Speech operated smart-home control system for users with special needs. In: 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, pp. 193–196 (2005)
3. Renals, S., Hain, T., Bourlard, H.: Recognition and interpretation of meetings: The AMI and AMIDA projects. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (2007)
4. Moeller, S., Krebber, J., Raake, A., Smeele, P., Rajman, M., Melichar, M., Pallotta, V., Tsakou, G., Kladis, B., Vovos, A., Hoonhout, J., Schuchardt, D., Fakotakis, N., Ganchev, T., Potamitis, I.: INSPIRE: Evaluation of a Smart-Home System for Infotainment Management and Device Control. In: Proc. LREC, pp. 1603–1606 (2004)
5. Hawley, M.S., Enderby, P., Green, P.D., Cunningham, S.P., Brownsell, S., Carmichael, J., Parker, M., Hatzis, A., O‘Neill, P., Palmer, R.: A speech-controlled environmental control system for people with severe dysarthria. Medical Engineering & Physics 29, 586–593 (2007)
6. Arking, R.: Biology of Aging. Oxford University Press, New York (2005)
7. Rabbitt, P., Anderson, M.: The lacunae of loss? Aging and the differentiation of cognitive abilities. In: Lifespan Cognition: Mechanisms of Change. Oxford University Press, New York (2006)
8. Deary, I.J., Whiteman, M.C., Starr, J.M., Whalley, L.J., Fox, H.C.: The impact of childhood intelligence on later life: Following up the Scottish Mental Surveys of 1932 and 1947. Journal of Personality and Social Psychology 86, 130–147 (2004)
9. Linville, S.E.: Vocal Aging. Singular Thomson Learning, San Diego (2001)

10. Ramig, L.O., Gray, S., Baker, K., Corbin-Lewis, K., Buder, E., Luschei, E., Coon, H., Smith, M.: The Aging Voice: A Review, Treatment Data and Familial and Genetic Perspectives. Clinical Linguistics and Phonetics 53, 252–265 (2001)
11. Xue, S.A., Hao, G.J.: Changes in the human vocal tract due to aging and the acoustic correlates of speech production: a pilot study. Journal of Speech, Language, and Hearing Research 46, 689–701 (2003)
12. Vipperla, R., Renals, S., Frankel, J.: Longitudinal study of ASR performance on ageing Voices. In: Proc.1 Interspeech 2008, pp. 2550–2553 (2008)
13. Baba, A., Yoshizawa, S., Yamada, M., Lee, A., Shikano, K.: Acoustic models of the elderly for large-vocabulary continuous speech recognition. Electronics and Communications in Japan, Part 2 (Electronics) 87, 49–57 (2004)
14. Wilpon, J.G., Jacobsen, C.N.: Study of speech recognition for children and the elderly. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 1, pp. 349–352 (1996)
15. Baeckman, L., Small, B.J., Wahlin, A.: Aging and Memory: Cognitive and Biological Perspectives. In: Handbook of the Psychology of Aging, pp. 349–377. Academic Press, San Diego (2001)
16. Verhaeghen, P.: Aging and vocabulary scores: a meta-analysis. Psychology of Aging 18, 332–339 (2003)
17. Shafto, M.A., Burke, D.M., Stamatakis, E.A., Tam, P.P., Tyler, L.K.: On the tip-of-the-tongue: neural correlates of increased word-finding failures in normal aging. J. Cogn. Neuro-sci. 19, 2060–2070 (2007)
18. Caruso, A.J., McClowry, M.T., Max, L.: Age-related effects on speech fluency. Seminars in Speech and Language 18, 171–179 (1997)
19. Pennebaker, J.W., Stone, L.D.: Words of wisdom: Language use over the life span. Journal of Personality and Social Psychology 85, 291–301 (2003)
20. Wolters, M., Georgila, K., Logie, R., MacPherson, S., Moore, J., Watson, M.: Reducing Working Memory Load in Spoken Dialogues: Do We Have to Limit the Number of Options? In: Interacting with Computers (accepted, 2009)
21. Barras, C., Geoffrois, E., Wu, Z., Liberman, M.: Transcriber: Development and use of a tool for assisting speech corpora production. Speech Communication 33 (2000)
22. Moore, J., Kronenthal, M., Ashby, S.: Guidelines for AMI Speech Transcriptions. AMI Deliverable (2005)
23. Carletta, J.: Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. Language Resources and Evaluation 41, 181–190 (2007)
24. Georgila, K., Wolters, M., Karaiskos, V., Kronenthal, M., Logie, R., Mayo, N., Moore, J., Watson, M.: A Fully Annotated Corpus for Studying the Effect of Cognitive Ageing on Users' Interactions with Spoken Dialogue Systems. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (2008)
25. Walker, M.A., Passonneau, R.J., Boland, J.E.: Quantitative and qualitative evaluation of DARPA Communicator spoken dialogue systems. In: Proceedings of the 39th Meeting of the Association for Computational Linguistics, pp. 515–522 (2001)
26. Anderson, S., Liberman, N., Bernstein, E., Foster, S., Cate, E., Levin, B., Hudson, R.: Recognition of Elderly Speech and Voice-Driven Document Retrieval. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Phonenix, Arizona (1999)
27. Vipperla, R., Renals, S., Frankel, J.: Longitudinal study of ASR performance on ageing voices. In: Proc. Interspeech, pp. 2550–2553 (2008)

28. Georgila, K., Wolters, M., Moore, J.: Simulating the Behaviour of Older versus Younger Users. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, Human Language Technologies (ACL/HLT), pp. 49–52 (2008)
29. Hain., T., Burget., L., Dines., J., Garau., G.: M.Karafiat., Lincoln., M., McCowan., I., Moore., D., Wan., V., Ordelman., R., Renals, S.: The 2005 AMI System for the transcription of Speech in Meetings. In: Proceedings of the Rich Transcription 2005 Spring Meeting Recognition Evaluation (2005)
30. Gauvain, J.-L., Lee, C.-H.: Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions on Speech and Audio Processing 2, 291–298 (1994)
31. Jurafsky, D., James, H.: Martin: Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. Prentice-Hall, Englewood Cliffs (2008)
32. Möller, S., Gödde, F., Wolters, M.: A Corpus Analysis of Spoken Smart-Home Interactions with Older Users. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (2008)