Spoken Dialogue Interfaces for Older People

Ravichander VIPPERLA [a,1], Maria WOLTERS [b] and Steve RENALS [b]

[a] *Multimedia Commmunications Group, Eurecom, France*

[b] *The Centre for Speech Technology Research, School of Informatics, University of Edinburgh, UK*

Abstract. Although speech is a highly natural mode of communication, building robust and usable speech-based interfaces is still a challenge, even if the target user group is restricted to younger users. When designing for older users, there are added complications due to cognitive, physiological, and anatomical ageing. Users may also find it difficult to adapt to the interaction style required by the speech interface. In this chapter, we summarise the work on spoken dialogue interfaces that was carried out during the MATCH project. After a brief overview of relevant aspects of ageing and previous work on spoken dialogue interfaces for older people, we summarise our work on managing spoken interactions (dialogue management), understanding older people's speech (speech recognition), and generating spoken messages that older people can understand (speech synthesis). We conclude with suggestions for design guidelines that have emerged from our work and suggest directions for future research.

Keywords. Spoken dialogue management, Automatic Speech Recognition, Speech synthesis

Introduction

Spoken communication with computing devices can offer a natural, intuitive and hands-free interface to the user. Speech-based systems have found increasing usage in information provision systems, software for personal assistance, in-car navigation systems, telecare, and in controlling home appliances. Spoken interfaces have the potential to be particularly useful for older people, especially those with mobility and visual impairments.

A typical spoken dialogue system (SDS) is shown in Fig. 1. The dialogue manager coordinates the other modules for speech input and output, and is usually provided with a knowledge base that contains the information specific to the task. In an SDS, spoken input from a user is processed by an automatic speech recognition module to generate a text transcript of the spoken command. If the system is designed to process complex spoken input, then a natural language understanding module is typically used to make sense of the user request. In order to communicate the response to the user, the dialogue manager invokes a natural language generation module that converts the response to a natural text format. This text is then converted into speech using a text-to-speech synthesizer.

[1] Corresponding Author: Ravichander Vipperla, Multimedia Communications Department, Eurecom, Sophia Antipolis, France ; E-mail: vipperla@eurecom.fr.

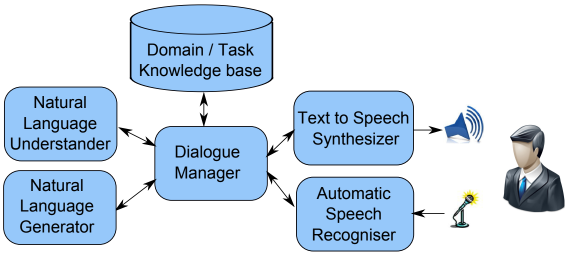

Domain / Task
Knowledge base
Natural Language Understander
Natural Language Generator
Dialogue Manager
Text to Speech Synthesizer
Automatic Speech Recogniser

Figure 1. Spoken dialogue system.

The use of an SDS in a home care setting can have a variety of uses such as medication reminders, home appliance control, on assistive robots or simply interactive chatbots. When designing and building such systems for older users, it is necessary to accommodate those changes that occur with ageing. In the MATCH project we focused on the development of SDS that takes account of cognitive ageing, auditory ageing, and vocal ageing—each of which has an impact of the usability of an SDS. In this chapter we outline the key findings of our research [1,2,3,4,5,6,7,8,9]. Following a review of the physiology of ageing with respect to spoken communication, we outline our research on how older people interact with dialogue systems. We also review our work on the two key speech technologies—speech recognition and speech synthesis—and how they should be adapted for older users.

1. Background

The impact of ageing is notoriously difficult to study because chronological age is a relatively poor predictor of anatomical, physiological, and cognitive changes [10,11]. This variability is not only genetic, but also depends on individual lifestyle [12]. As a consequence, designing systems for older users is challenging because individual older people have very different needs and abilities.

Ageing mostly affects aspects of cognitive function such as reasoning [13] or working memory [14]. Acquired knowledge such as vocabulary tends to be well preserved [15]; older users may even have a richer vocabulary than younger ones. However, older users are more prone to difficulties in finding words and may produce more disfluencies in stressful conditions [16].

Ageing also affects the physiology of speech and hearing. The human vocal mechanism consists of the lungs, the larynx (which houses the vocal cords), and the vocal tract comprised of the pharynx, the mouth and the nose. Typical changes associated with ageing are reduction in the respiratory muscle strength and lung elasticity that result in lower energy in produced speech, restricted vocal fold adjustments during phonation that affect voice quality, and difficulty in producing the required tongue and lip shapes [17], which can affect articulatory patterns and speech rate.

Speech perception is also affected by age. The most well-known effect of auditory ageing is that low intensity sounds, in particular those at high frequencies, become more difficult to hear [18]. For example, older people find it more difficult to understand rapid or time-compressed speech [19], and speech in reverberant [20] or noisy environments [21,22].

Speech interfaces for older people have been developed in a variety of contexts, ranging from systems that are intended to cover the whole population, such as airline help systems, to specialist health care and rehabilitation applications such as symptom management (e.g.[23,24]), delivery of health care interventions (e.g. [25]), memory aids (e.g. [26]), home-based assessment using kiosks (e.g. [27]), and falls monitoring (e.g.[28]).

Dulude [29] evaluated the usability of six commercially deployed Interactive Voice Response systems with 22 younger and and 22 older people Only one of these systems used speech input; the others relied on a touch-tone setup. The performance of older users varied greatly. Even though older and younger users experienced similar problems with the systems, older users' performance suffered more, and they were more likely to give up than younger users. Older people were particularly affected by voices that spoke too fast, wrong keystrokes, and missing opportunities for error recovery.

Zajicek proposed a dedicated set of design patterns for voice interfaces tailored to older users [30]. One example is *Menu Choice*, which suggests that menus should be limited to at most three items. These design patterns are mainly motivated by existing HCI guidelines and results from cognitive psychology.

2. Dialogue Management

In the MATCH project, we focussed on understanding how older people interact with dialogue systems. In particular, we tested two strategies for reducing working memory load, and we examined whether there are consistent differences between the interaction style of younger and older users that can be modelled using simulated users.

2.1. Experimental methodology

We tested two strategies for reducing working memory load [1]: reducing the number of options (four alternatives, two alternatives, yes/no); and confirming agreed information (explicit confirmation dialogues, implicit confirmations where agreed information is merely repeated, no confirmations). The strategies are illustrated in Table 1.

Table 1. Strategies for Reducing Working Memory Load

Options (Health Professional)		**Confirmation (Physiotherapist)**	
	System		System
Yes/No:	Would you like to see the physiotherapist?	**Explicit:**	You would like to see the physiotherapist. Is that correct?
2 Options:	Would you like to see the physiotherapist or the diabetes nurse?	**Implicit:**	When would you like to see the physiotherapist, on Thursday afternoon or on Friday morning?
4 Options:	Would you like to see the occupational therapist, the community nurse, the physiotherapist or the diabetes nurse?	**None:**	*System:* When would you like to come, on Thursday afternoon or on Friday morning?

We recruited 50 participants, of whom 26 were older (aged 50–85) and 24 were younger (aged 18–30). The older users contributed 232 dialogues, the younger ones 215. Older and younger users were matched for level of education and gender. These participants completed a series of four tests to assess two main dimensions of intelligence: fluid intelligence, which is linked to abstract reasoning, and crystallised intelligence, which is linked to acquired knowledge as well as working memory and information processing speed. All tests were presented visually, to avoid problems due to age-related hearing loss [31].

Fluid intelligence was assessed using Ravens' Progressive Matrices [32]. *Crystallised intelligence* was measured using the Mill Hill Vocabulary test [32]. *Working memory capacity* was assessed using a sentence reading span test implemented in ePrime [33]. *Information Processing Speed* was assessed using the Digit Symbol Substitution subtest (DSST) of the Wechsler Adult Intelligence Scale-Revised [34].

Overall, we found that crystallised intelligence increases with age, and information processing speed, fluid intelligence and working memory capacity decrease with age.

To explore the dependence of dialogue management strategies with age, the experimental participants performed an appointment scheduling task, by interacting with nine different SDSs, one for each combination of strategies. Each SDS was simulated using a Wizard-of-Oz approach, where the "wizard" was responsible for speech recognition and natural language understanding, while dialogue management and natural language output generation was done automatically. Each system used a strict system-initiative policy. In the experiments, users first arranged to see a specific health care professional, then they agreed on a specific half-day, and finally, a specific half-hour time slot on that half-day. Users were unable to skip any stage of the dialogue. This design ensured that all users were presented with the relevant number of options and the relevant confirmation strategy at least three times per dialogue. At the end, the appointment was confirmed. After interacting with a SDS, participants filled in a detailed usability questionnaire. This yielded both user satisfaction data and served as a brief distraction. Having completed the questionnaire, they were asked to recall the appointment. Participants could not take notes and had to rely on memory.

Both younger and older users correctly scheduled their appointments and were able to recall appointments well regardless of dialogue strategy. This is contrary to the design recommendations of e.g. Zajicek [30]. One possible reason for this result is that our task could be solved by waiting for the first acceptable option, which removes the need to remember, compare, and contrast options. In this case, presenting more options at a time is better, especially for users with low working memory capacity [35].

An examination of dialogue length showed that using confirmations makes older users less efficient. When the system used implicit confirmations or no confirmation at all, older users' dialogues were only around two turns longer than younger users' dialogues. When the system used explicit confirmations, on the other hand, the gap widened to around five turns.

The overall impression of the system was not affected by dialogue strategy. On the whole, older users were less satisfied with the systems than younger users. For more details on the experiment design and statistical analysis, see [1].

2.2. Experimental analysis

The dialogues that were generated during the study were recorded, transcribed orthographically, and annotated with dialogue acts and dialogue context information, to form the MATCH dialogue corpus [36]. Using a unique mapping, we associate each dialogue act with a ⟨speech act, task⟩ pair, where the speech act is task independent and the task corresponds to the slot in focus (health professional, half-day or time slot). For example, ⟨confirm pos, hp⟩ corresponds to positive explicit confirmation of the health professional slot. For each dialogue, detailed measures of dialogue quality were recorded: objective task completion (appointment recall), perceived task completion, length (in turns), and user satisfaction ratings as provided in the questionnaire.

Based on this corpus, we compared the interaction style of younger and older users. In particular we investigated if users be categorised into distinct groups depending on how they speak to the system, and how such groups might be characterized, and whether the interaction style of an individual cab be predicted.

The interaction style of each user was characterised by features such as dialogue length, frequencies of speech acts, and word frequency features. In order to establish the underlying cluster structure, we used hierarchical agglomerative clustering and partitioning based top-down clustering. Classes were assigned based on agreement. A strong pattern of two dominant clusters emerged. The clusters can be described by two key words, social and factual. Factual users adapted quickly to the SDS and interacted efficiently. Social users, on the other hand, treated the system like a human, failing to adapt to the SDS's system-initiative dialogue strategy. While almost all the social users were older, about a third of the factual users were found to be from the older age group. The results suggest that while older users are more social, it is advisable to adapt the spoken dialogue systems to users based on their observed behaviour rather than simply relying on their age.

2.3. Learning dialogue management

A dialogue manager can be based on hand-crafted rules or on statistical models obtained using reinforcement learning [37,38]. In the statistical approach, strategies that lead to successful dialogues are rewarded, while strategies which lead to failure are penalised. Reinforcement learning requires a substantial number of recorded dialogues, which is currently difficult to obtain from human users. Therefore, approaches have been developed using simulated users [39] in which the number of dialogues is limited only by computing power. Although simulated users must be learned from interactions between systems and human users, the number of interactions needed to learn a good simulated user is several orders of magnitude lower than the number of interactions needed to train a statistical dialogue management model.

Using the MATCH corpus, we constructed the first known age-sensitive user simulations [3]. We showed that simulations based on younger users could not appropriately describe the behaviour of older users, but that simulations created on the basis of older users or all users covered behaviour patterns of both younger and older users fairly well.

We tested the utility of these user simulations for learning dialogue policies in two further experiments [4]. In the first experiment, policies were rewarded that filled slots in the correct order (strict policy). A large penalty was imposed when the policy deviated from the strict slot order (health professional, half-day, time slot). In the second experiment, these constraints were removed and slots could be filled in any order (user-initiative policy). For each experiment, two policies were learnt, Policy-Old, which was based on simulated older users, and Policy-Young, which was based on simulated younger users. The resulting policies were then tested on simulated older users (Test-Old) and simulated younger users (Test-Young). Scores for each combination of policy and simulated user were established using 5-fold cross-validation.

The strict policy that was learned from simulated younger users was as follows, with only slight variations: first request the type of health professional, then implicitly confirm the health professional and request the half-day slot, then implicitly confirm the half-day slot and request the time slot, and then confirm the appointment.

The flexible policy learned from simulated older users was a much better fit for their interaction style than the strict policy. The score for the flexible policy learned from simulated younger users was relatively low, even though the resulting policy was very similar to the strict policy learned from younger users (i.e. a sequence of information requests and implicit confirmations), and even though the behaviour of younger users is far more predictable than the behaviour of older users.

The results indicate that simulated users can be used to learn appropriate policies for older adults, even though their interaction behaviour is more complex and diverse than that of younger adults. Crucially, simulated older users made it possible to learn a more flexible version of the strict system-initiative dialogue strategy.

3. Speech Input

In this section we present some automatic speech recognition experiments, comparing older and younger users, and investigate the acoustic changes that arise from ageing, and their effects on speech recognition accuracy.

We first present the basic architecture of automatic speech recognition (ASR) systems that convert speech input to text which can subsequently be processed by the dialogue system. We then summarize our work on understanding the impact of several changes observed with vocal ageing on ASR accuracies.

State-of-the-art ASR systems recognise speech by predicting the most likely sequence of words that can explain an observed sequence of recorded acoustic information [40]. Such an approach is based on two statistical models estimated from data: the acoustic model, which provides an estimate of the likelihood of an acoustic observation sequence given a word sequence; and the language model which provides a prior probability for a word sequence. A feature extraction module assumes the task of converting the speech signal into discrete parameter vectors, and the lexicon acts as a map between the words in the language model and the sub word units (typically phones) that comprise the acoustic models. Given some recorded speech (acoustic observations), the most likely sequence of words is found using a search, or decoding process based on the acoustic and language models. Figure 2 illustrates such a system.

The acoustics are typically modeled with statistical models such as Gaussian mixture models (GMMs) sequenced using hidden Markov models (HMMs). All the results presented here have been conducted in this framework, with the front end parametrization being perceptual linear prediction coefficients (PLP) which are one of the best known

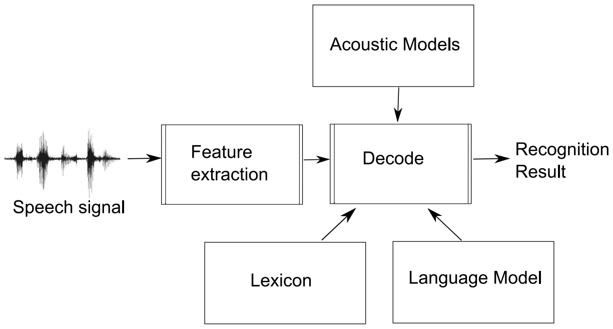


Figure 2. Automatic speech recognition system

features for speech recognition purposes. The ASR results are presented in terms of word error rate (WER) which is a combination of substitution, deletion and insertion errors in the recognition hypothesis with reference to the ground truth transcription.

Our work in the MATCH project [41,5,6,42,7] was primarily focused on understanding how ageing impacts the ASR accuracies from an acoustic modeling point of view. Four different corpora were used in our experiments: The supreme court of United states corpus (SCOTUS); The Japanese News paper archive sentences (JNAS and senior JNAS) corpus; The Augmented multiparty interaction corpus (AMI); and the MATCH corpus. While most analysis results were validated on at least two different corpora [7], the results on only one corpus for each analysis are presented here for brevity.

In order to understand how ASR word error rates vary between younger and older users, experiments were performed on the SCOTUS corpus, the details of which can be found in [41]. The test set comprised 27 younger speakers (23 Male and 4 Female) in the age range of 30-45 years, and 12 older speakers (10 Male and 2 Female) in the 60-85 years age range. The results presented in the Figure 3 show a significant difference of 9.3% absolute higher WERs for older speakers compared to younger speakers. Standard approaches for improving accuracies such as adapting the speaker independent models to the target speaker using maximum likelihood linear regression (MLLR), vocal tract length normalisation (VTLN) to compensate for the differences between speakers and the speaker adaptive training (SAT) that seperates speaker specific characteristics from the phone acoustic models [43] also fail to bridge the gap in WERs for the two age groups.

In order to understand the reasons behind these differences, we analysed the effect of glottal source characteristics and articulatory changes.

3.1. Glottal source characteristics

Some of the most perceptually apparent vocal changes that occur due to ageing, such as increased hoarseness, are due to the changes in the glottal source characteristics. These changes are characterised by changes in parameters such as fundamental frequency, jitter, shimmer, and the noise-to-harmonic ratio (NHR). Jitter and shimmer measure the perturbations in time and amplitude of a signal (Fig. 4). Jitter is the ratio of pitch period (T_i) variation from cycle to cycle to the average pitch period, expressed as a percentage. Shimmer is the average absolute difference between the amplitudes (A_i) of consecutive

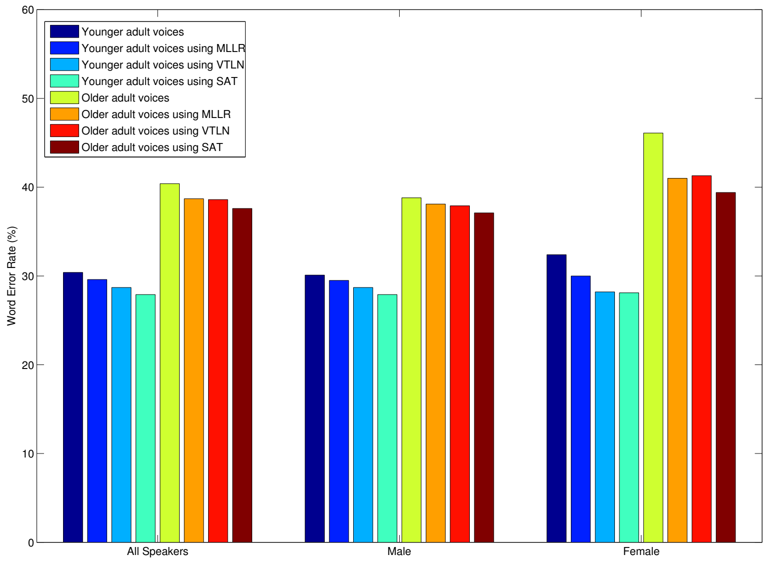


Figure 3. Comparison of ASR WERs between younger and older adults in the SCOTUS corpus

periods, divided by the average amplitude, also expressed as a percentage. The noise-to-harmonic ratio (NHR), which correlates well with breathiness in the voice, is the ratio of the noise to the energy in the periodic part of the signal.

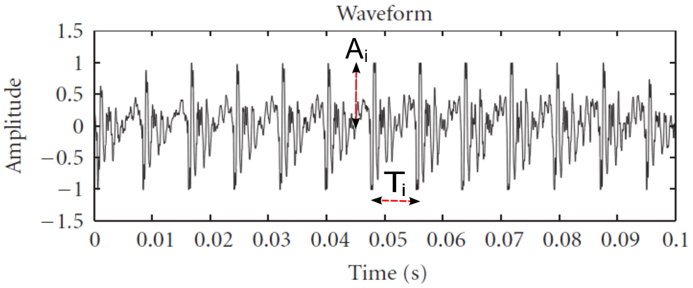


Figure 4. Perturbations in time and amplitude lead to jitter and shimmer respectively

Voice analysis is typically carried out on sustained vowel pronunciations recorded in a controlled noise-free environment. However the SCOTUS corpus is spontaneous speech with a considerable amount of background noise. Being spontaneous in nature, the corpus also does not have sustained vowel pronunciations, with most of vowel samples being a fraction of a second long and part of a longer utterance. In order to pick the best available instances of the sustained vowel (*aa*) phonations from the speech, each utterance from the male speakers in the test set was aligned to the phonetic transcription in order to determine the phone boundaries and the likelihood of the phone sequence

Voice quality measures	Younger Males		Older Males		p-value
	Mean	Std	Mean	Std	
Mean F0	143.9	43.2	128.0	44.6	< 0.001
Jitter (Local)	1.89	1.50	2.41	1.83	< 0.001
Shimmer (Local)	10.73	5.22	11.33	5.27	< 0.001
NHR	0.21	0.15	0.21	0.16	0.79

Table 2. Comparison of voice parameters between younger and older male speakers for the phonations of vowel *aa*

comprising the utterance. Segments of the utterance corresponding to the phone *aa* satisfying a minimum length of 0.1 seconds and a likelihood threshold criterion were chosen for voice analysis. In all, 2970 samples from 23 adult male speakers and 2105 samples from 10 older male speakers were used for voice analysis. The analysis was carried out only on male speakers to avoid any gender related confounding effect in the results.

The comparative results of the differences in these parameters between younger and older males are displayed in Table 2. We observe that F0 reduces by about 10% and there is a significant increase in jitter and shimmer measures. In order to understand if these changes contribute to the increased WERs in older voices, a set of experiments were conducted to artificially introduce these artifacts into younger adult voices [6].

To understand the F0 impact, the frequencies were scaled to 0.9 of their original value to reflect the values observed in older voices [6]. The word error rates before and after reduction in $F0$ are given in Table 3. The WER increased by 1.1% absolute to 33.2% and is statistically significant with $p < 0.001$.However, this difference could be overcome to a certain extent by using vocal tract length normalisation.

Word Error Rate (WER) %			
	Original	Reduced $F0$	p-value
Without VTLN	32.1	33.2	< 0.001
with VTLN	28.8	29.5	< 0.01

Table 3. WER (%) with artificial reduction in fundamental frequency of the speech from younger adults in the SCOTUS corpus.

Jitter and shimmer in the younger adult test set were manipulated similarly, with a maximum allowable perturbation of 5% and 10% ($\alpha = 0.05$ and 0.1 respectively). ASR results are shown in Table 4, where it can be seen that the changes in glottal source characteristics have minimal impact on the ASR WERs. Front end feature extraction techniques in ASR such as perceptual linear prediction used in our experiments are designed to capture the vocal tract characteristics and surpress the glottal characteristics. As a result, the changes in glottis with vocal ageing has little impact on ASR accuracies.

3.2. Articulatory changes

While there are significant differences in the glottal source characteristics, they do not explain the differences in WER for the the two age groups. ASR systems model articulatory patterns through the acoustic model. We investigated three aspects of articulatory changes: phone recognition accuracy, vowel centralisation, and speaking rate.

Word Error Rate (WER) %			
Artifact introduced	**Original**	**Modified**	
		$\alpha = 0.05$	$\alpha = 0.1$
Increase Jitter	32.1	32.2 ($p = 0.62$)	32.4 ($p = 0.17$)
Increase Shimmer	32.1	32.1 ($p = 0.65$)	32.1 ($p = 0.13$)

Table 4. WER (%) with artificial modification of the speech from younger adults in the SCOTUS corpus to reflect the voice parameter values observed in older adults

In order to understand which phones are most impacted by ageing in terms of ASR WERs, we performed an experiment on the SCOTUS corpus using the same train and test set as for the baseline experiments above, using a phone loop decoder (where any phone is allowed to follow any phone in an unconstrained manner) to output a phone transcription purely based on the acoustic model likelihood. In such a setting, the confounding effect of the language model are eliminated and an insight into the acoustic differences becomes possible. The results in this case are measured using phone correct recognition percentage which is a ratio of the number of occurences of a phone in the decoded hypothesis to the total number of occurences of the phone in the reference hypothesis. These results indicated that the class of phones whose recognition accuracy decreased the most with ageing were vowels [7].

We hypothesised that this impact is due to vowel centralisation, in which the discriminability between vowels is reduced due to articulatory undershoot, leading to changes in vowel formant frequencies towards the mean. It has been reported in [44] that vowel centralisation is quite pronounced in very old speakers with all vowel realisations sounding quite close to each other. Vowel centralisation is typically measured using the vowel space area. First and second formant frequencies (F1 and F2) are calculated for each vowel and the vowels are plotted in the 2 dimensional F1-F2 space. The vowel space area is the area enclosed by the corner vowels *i, u, a* and *ae*.

The vowel space analysis was carried out on the utterances used for the experiments on glottal source characteristics described previously. The values of first (F1) and second formant (F2) frequency for each vowel instance were computed at the vowel midpoint. After rejecting outliers, mean values of each vowel for a speaker were computed.

The vowel space bounded by the phones *aa, uw, iy* and *ae* for both the age groups is shown in figure 5. The corner points of each quadrilateral is the average across all the speakers in that age group. The area of the vowel quadrilateral for each speaker was computed, and shown in Table 5. The area occupied by the vowel quadrilateral of older speakers is less than that of younger speakers, indicating vowel centralization. The vowel space areas of the speakers in the two age groups are significantly different at $p < 0.01$ using Student's T test.

Vowel space area (Hz^2)			
Younger adult males	**Older adult males**	**Difference**	**p-value**
5.46×10^4 (std: 1.33×10^4)	3.99×10^4 (std: 0.97×10^4)	1.47×10^4	< 0.01

Table 5. Vowel Space Area comparison between *younger adult* and *older adult* males in SCOTUS corpus

While all the corner vowels appear to shift in the F1-F2 space with ageing, the phone recognition results indicated that vowels *ae* and *aa* have a large decrease in performance

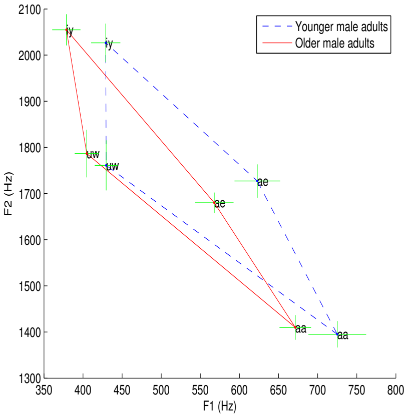
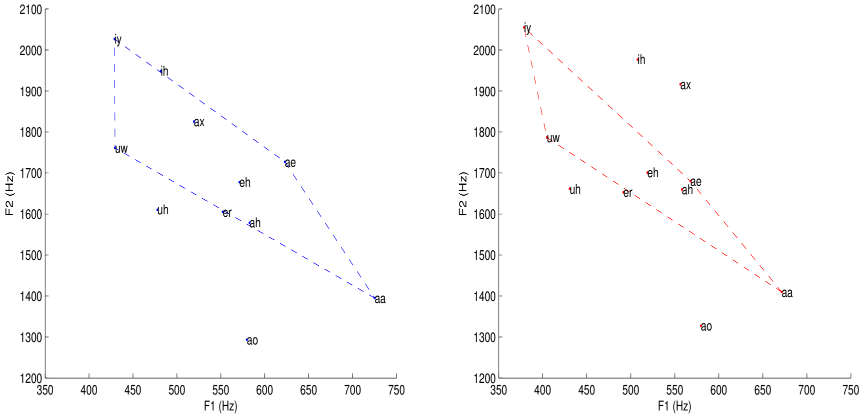


Figure 5. Mean vowel space areas for younger and older male adults in the SCOTUS corpus. Corner vowels and their standard deviations are also shown in the figure.



(a) Vowel space: younger adults

(b) Vowel space: older adults

Figure 6. Centroid positions of common vowels in younger and older Adults

with ageing while the phones *iy* and *uw* do not show much difference in accuracies. In order to understand this better, the centroids of all the vowels (averaged over all the speakers in each age group) are shown in Figure 6. For the older speakers, the vowels appear to move closer to each other into clusters in the F1-F2 space especially in the central region of the vowel space. This might lead to a reduction in the discrimination capacity between phones and also explain to some extent the large decrease in recognition accuracies for some of the central vowels.

The third articulatory pattern we explored, speaking rate, has been reported to be slower in older adults compared to younger adults. Though the underlying physiological change for the decrease in speech rate is not clear completely, it is believed to be a result of restricted free movement in articulators and a reduction in motor control capabilities.

It has also been suggested that older adults tend to deliberately reduce speech rate so as to be more intelligible under restricted motor control abilities. Slower speaking rate is a combined effect of longer pronunciation of words, increased number of pauses and pause duration. The impact of speaking rate differences on ASR accuracies has received little attention in ASR research.

For our analysis of speaking rate, we computed the average duration of each phone. We used the JNAS corpus for the analysis where the test set comprised 101 older and 101 younger speakers with about 50 utterances per speaker, with the speakers in the two age groups being roughly balanced in terms of gender. The average duration (d_p) for each phone for each age group was computed and a weighted average phone duration (based on the phone frequency) for older and younger speakers is shown in Table 6. The results show a statistically significant decrease ($p < 0.01$) in the speaking rates of older speakers.

Speaking Rate (msec per phone)			
Younger Males	**Older Males**	**Younger Females**	**Older Females**
72.0	91.2	78.1	93.8

Table 6. Speaking rate differences between younger and older adults in the JNAS corpus

In an experiment to understand the impact of using acoustic models trained on speech with a speaking rate different from that of the test set, seperate acoustic models were trained on the JNAS corpus for each age-gender group [7]. The state transition probabilities of these acoustic models capture the duration information for each phone and thereby speaking rate. A different set of acoustic models were then created by swapping the transition probabilities of the models of older males with those of younger males and similarly for the female speakers. All other parameters of the original models were kept fixed and the test sets decoded with the original and modified models. The results shown in Table 7 capture in effect, the outcome of slower speech test set decoded on models trained on slower speech and models trained on faster speech, all other characteristics of training speech being the same. It is observed that while correct recognitions are almost the same, the word error rates increase for both male and female speakers with modified models. It can thus be concluded that insertions errors increase for slower speech decoded with models trained on relatively faster speech.

Acoustic models	**Older males**		**Older females**	
	% Correct	**% Accuracy**	**% Correct**	**% Accuracy**
Original	75.8	69.6	84.3	80.0
Modified	75.6	68.1	84.2	79.6

Table 7. Word correct recognition and accuracies for older speakers in the JNAS corpus with original and transition parameter modified models

3.3. Acoustic modelling for older speakers

One of the questions of interest is to address whether speech from older speakers is different from that of younger speakers in the acoustic feature space used in ASR. In order to answer this, a speaker-age group classification task was set up on the MATCH corpus.

Using speaker independent acoustic models trained on the AMI corpus, speaker adaptation transforms for each of the 50 speakers in the MATCH corpus were computed. These transformation matrices capture speaker specific information and can be used as speaker identities. The similarity between every pair of speaker transforms was computed [42] and using those similarity scores all the speakers were then clustered into four groups using CLUTO [45] by the repeated bisections method. In this method, the speakers are first clustered into 2 groups, which are bisected again to obtain four groups.

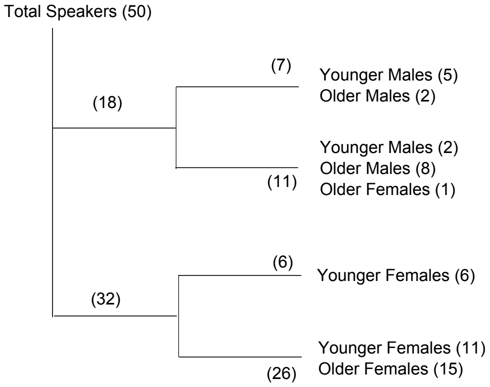


Figure 7. Clustering of speakers in the MATCH corpus

The speaker distribution in the clusters is shown in Figure 7. As expected, at the first level of clustering, the male and female speakers are separated out. At the next level of clustering, there appears to be some separation between the younger and older male speakers. While for the females, there is a large overlap of younger and older females in a cluster. This also corroborates with the fact that age related changes in the voice for women are less pronounced than those observed in men.

Motivated by the results of speaker clustering, hierarchical acoustic models were built on the JNAS corpus as shown in figure 8 to find if any improvements in ASR WERs were possible for the older age group. The root speaker independent models were trained using the entire training set of 205 younger speakers and 187 older speakers roughly balanced in terms of gender. The acoustic models in each child node are derived from the parent node by adapting with a node specific subset of the training set using maximum a-posterori approach.

The results shown in Table 8 are obtained by decoding age and gender dependent subsets of the test set with various sets of acoustic models in the hierarchical structure. While gender dependent models achieve significant improvements over speaker-independent models, there are modest additional gains possible with further clustering based on age for the older age group. Such age based grouping is observed to be counter productive for younger adult group and they appear to benefit from a larger training set (including speech from older speakers).

It can thus be concluded that while it helps to a certain extent to collect more data from the older age group to improve their ASR accuracies, the accuracies that can be achieved are still lower than what is desired. Perhaps speaker specific models where the

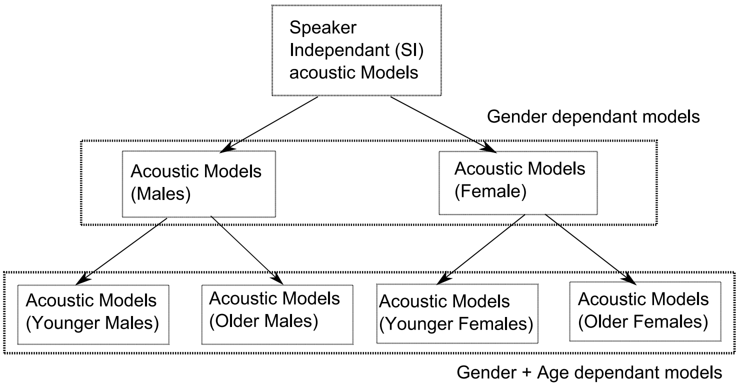

Speaker Independant (SI) acoustic Models
Gender dependant models
Acoustic Models (Males)
Acoustic Models (Female)
Acoustic Models (Younger Males)
Acoustic Models (Older Males)
Acoustic Models (Younger Females)
Acoustic Models (Older Females)
Gender + Age dependant models

Figure 8. Gender and age dependent acoustic models

ASR Word Error Rates (%)				
Acoustic models	**Younger Adults**		**Older adults**	
	Male	**Female**	**Male**	**Female**
Speaker Independent	16.2	15.5	23.9	16.9
Gender Dependent	15.1	14.3	21.4	15.9
Gender and Age Dependent	15.7	14.2	21.1	15.7

Table 8. ASR WERs using hierarchical models based on gender and age groups

changes in the articulatory patterns for each individual are carefully taken into consideration during modeling might form an appropriate solution. Building a framework for such modeling is a subject of future work in this direction.

4. Speech Output

Computer-generated synthetic speech makes it easy to adapt the spoken messages used in home and telecare systems to users' needs and preferences. Full speech synthesis systems can be used to generate virtually any sentence desired. This means that content and wording can be adapted quickly and cost-effectively to changing care needs. Designers can choose from a range of voices, speaking styles (and, in the case of avatars, faces) to provide users with a choice of personas for their interaction with the system.

Despite recent advances in speech synthesis technology, synthetic speech can still be seen as degraded relative to human speech because the synthesis process may introduce artefacts, fail to include potentially important information in the signal or even produce mistakes. In the light of the effect of ageing on hearing, we need to ask what effect this degradation has on older users' ability to understand synthetic speech.

Formant synthesis, the technology used in DECTalk [46], is perhaps the most artificial of all forms of speech synthesis. In formant-based systems, the speech signal is created from scratch based on a model of the speech production process. It therefore lacks much of the rich acoustic information found in natural speech. Due to its long tradition and its widespread use in augmentative and alternative communication, formant synthesis is by far the best studied synthesis technology for older people. Using a task involving the recognition of isolated monosyllabic words, the Modified Rhyme Test [47], Humes [48] showed that older people with a hearing loss can understand formant synthesis as well as natural speech. However, when moving from words to sentences, new problems emerge. The dearth of acoustic information in the signal [49] and wrong intonation patterns [50] may make it more difficult for the listener to understand what is being said, which in turn increases cognitive load [51]. Due to reduced cognitive resources, a higher cognitive load may affect older listeners more than younger ones [52]. Some strategies that increase intelligibility of formant synthesis are reducing the speech rate [53] or making sure that appropriate context is provided [54].

Roring, Hines, and Charness [55] studied another form of speech synthesis, diphone synthesis. In diphone synthesis, new utterances are constructed from units that are designed around the transition between two speech sounds. Roring *et al.* asked younger and older listeners to identify monosyllabic words that were either spoken in isolation or occurred in the final position of a sentence that provided a few contextual cues. Overall, older listeners found synthetic speech more difficult to understand, even when the material was slowed down by 50%. Older and younger listeners profited equally from additional context. All age-related effects could be explained by increased pure-tone thresholds. Even though they performed thorough cognitive assessments, Roring *et al.* did not observe any additional effect of cognitive ageing.

4.1. Unit selection speech synthesis

During the MATCH project, we focused on unit selection synthetic speech [56], which is both highly intelligible and sounds very natural. Unit selection systems generate output by concatenating segments of speech that are taken from a large database of pre-recorded utterances. These segments can be as small as the transition between two sounds used in diphone synthesis, or as large as entire phrases. Given input text, the unit selection algorithm searches the database for the best-fitting sequence of segments. Unit selection databases typically contain about ten hours of high-quality speech from a variety of genres, recorded by a single speaker.

Unit selection synthesis is characterised by requiring very little additional signal processing that might distort the speech. The acoustic richness of the original human voice is preserved, and the resulting output often sounds highly natural. The downside is that modifying the intonation, speaking rate, and rhythm of the resulting speech may be difficult. As most unit selection systems are highly modular, it is relatively easy to add new voices, genders, and dialects. In later work, not described here, we have carried out intensive research on statistical parametric speech synthesis, an approach which has the advantages of offering much greater control over the generated speech, as well as offering the possibility to personalise the synthetic speech through speaker adaptation, an approach which requires only a few minutes of speech data [57]. This technology has been applied to voice reconstruction for people with degenerative diseases such as motor neurone disease and Parkinson's disease [58], and was compared to unit selection speech synthesis in terms of intelligibility in experiments performed using Amazon Mechanical Turk [59].

Lines and Hone investigated the intelligibility and acceptability of an early version of unit selection, which used units of flexible length that can be up to a word long [60], with older users [61,62]. The target application was an interactive domestic alarm system

	Younger Users		Older Users	
Number	20		42	
Age in Years	M=25	SD=4.4	M=60	SD=7.4
Perc. Male	N=4	20	N=18	43
Difficulty Hearing	N=1	5	N=24	57
Hearing Loss, at least one ear				
Based on PTA4	N=0	0	N=17	40
Conductive	N=0	0	N=7	17
Mixed	N=0	0	N=4	10
Hearing Loss, bilateral				
Based on PTA4	N=0	0	N=17	40
Conductive	N=0	0	N=7	17
Mixed	N=0	0	N=4	10

Table 9. Participant Demographics

installed in four trial homes of the Millenium Home project [28]. They compared a male and female synthetic voice to a male and female human voice. All voices spoke British English with a Received Pronunciation (RP) accent. To assess intelligibility, users were asked to perform a series of tasks around the home in each of the four voices. Task success was uniformly high for each voice. However, the synthetic voices were seen as less natural, less pleasant, more irritating, and more boring.

Langner and Black [63] evaluated fully-fledged unit selection where segments were selected from a large database and were allowed to be longer than a word. Younger and older participants received computer-generated information about bus timetables over a telephone. Participants transcribed what they heard, and the word error rate of these transcriptions was measured, with the finding that people with self-reported hearing problems found it more difficult to understand messages spoken by a computer voice.

4.2. *Experimental study*

Since formant synthesis and unit selection synthesis are so different, it is not straightforward to translate message design principles developed using formant synthesis systems to unit selection. While it is possible to slow the down the output of formant synthesisers by almost arbitrary amounts, this is not so easy for unit selection. The signal processing methods that are used to stretch unit selection speech can introduce undesirable artefacts if the speaking rate is changed by more than 10-20%. In this study, we investigated the baseline intelligibility of unit selection systems with older listeners under good listening conditions with a view to proposing a new set of guidelines that are largely independent of the synthesis technology used.

The demographics of the participants in terms of ageing and hearing loss are summarised in Table 9. Relevant medical and family history information was obtained using a detailed questionnaire. Working memory capacity was assessed using a visually presented reading span test [64]. Visual evidence of ear disease was screened for using otoscopy. The full set of tests, which is described in detail in [8], included pure-tone audiometry, bone-conduction audiometry tympanometry, speech audiometry in quiet [65], and a gap detection test [66].

Severity of hearing loss was described based on the average air-conduction thresholds at 0.5, 1, 2, and 4 kHz (PTA4) following the guidelines of the British Society of Audiology [67]. Hearing was considered normal when PTA4 air-conduction thresholds were $\leq$ 20 dB(HL) in both ears. People with a PTA4 air-conduction threshold $>$20 dB(HL) were considered to have a hearing loss. A PTA4 threshold of between 20 and 40 dB(HL) indicated mild hearing loss, a threshold of between 40 and 70 dB (HL) indicated moderate hearing loss, and a threshold of 71 dB or higher indicated severe hearing loss. Following BSA guidelines, a bone conduction threshold that was at least 10 dB lower than the corresponding air-conduction threshold was taken to indicate a potential conductive hearing loss, i.e. problems with hearing due to middle ear pathology. An air-bone gap of at least 10 dB together with an average bone-conduction threshold of 20 dB (HL) or higher was taken to indicate a mixed hearing loss.

The speech task covered two types of reminders that are particularly relevant to the home care domain: reminders to meet a specific person at a given time (*appointment reminders*), and reminders to take a specific medication at a given time (*medication reminders*). Participants heard a reminder and were asked to recall either the time or the other relevant information provider (person / medication).

We implemented a full factorial design with five variables, reminder type (appointment / medication), reminder voice (human/synthetic), order of information (time first / time last), information to be recalled (time / person or medication), and position of information to be recalled (non-reminder-final / reminder-final) for a total of $2 \times 2 \times 2 \times 2 \times 2 = 32$ reminders. Only the voice of the reminder itself was manipulated; questions were always presented using the human voice to ensure participants could understand the recall task.

Times are remembered well regardless of voice or position in the sentence, at 95% correct. Person names are remembered less well when they are presented using the synthetic voice (86%) instead of the human voice (98%). Medications were by far the most difficult response category. Overall, people only remembered half the medication names correctly (50% correct). It was only for the combination of this particular response category and the synthetic voice that we observed a recency effect. While people only remembered one in three medication names correctly (30% correct) when they occurred in the first slot in a synthetic reminder, they were able to remember two in five when the name occurred in the second, final slot (44% correct). Multilevel modelling confirmed the findings of Roring et al. [55] that most of the age-related variation was due to differences in hearing, not cognition.

5. Conclusions

Designing spoken dialogue systems for older people presents many challenges. The studies reported in this chapter have shown that simple systems with basic functionality can potentially work well for younger and older people. When creating new spoken dialogue interfaces for home care tasks, we suggest the following practical guidelines:

Speech Recognition and Natural Language Understanding: If the speech recognition component used does not work as well for older as for younger people, create materials that reflect the speaking style of the older user who will be interacting with the system and schedule a session for training and adapting the recogniser.

These materials should cover dialectal variation as well as typical sentence structures and expressions.

Dialogue Management: Adapt to the user's interaction style. Are they quite chatty or are they likely to give short, focused answers? Ensure that the dialogue structure follows existing mental models, like the ones discussed here for medication. When adapting to cognitive limitations, be wary of making assumptions based on the literature. A detailed cognitive walkthrough will show you what information needs to be processed and memorised.

Speech Synthesis and Natural Language Generation: Speak clearly, using the appropriate settings for the speech synthesis system you are using. Use of phrasing and, if available, intonation to highlight important information can be achieved in different ways depending on the system. Sometimes intelligibility can be improved by slowing down the generated speech, but rate of slowing depends on system.

The biggest future challenge will be to assess the effect of speech recognition problems on the usability of spoken dialogue systems, so that strategies can be developed to address them. There are also very few studies of older people using deployed speech interfaces, with the notable exception of [29]. This is a general problem in spoken dialogue systems research, which is mostly conducted in the lab, not in the wild.

We also need a better understanding of the advantages and limitations of spoken dialogue systems in order to decide where and when to use them. In related user requirements work [68], we have found that many users prefer visual interfaces. Even among those who would use auditory interfaces, there is a divide between people who prefer speech and people who prefer non-speech audio. While ongoing work in mainstream speech technology and dialogue research is focusing on increasing robustness and usability, work within the Human-Computer Interaction field on speech and dialogue technology for older adults should also concentrate on identifying relevant use cases and improve screening for usability problems.

Acknowledgements

References

[1] M. Wolters, K. Georgila, R. Logie, S. MacPherson, J. D. Moore, and M. Watson. Reducing working memory load in spoken dialogue systems. *Interacting with Computers*, 21(4):276–287, 2009.

[2] M. Wolters, K. Georgila, S. MacPherson, and J. D. Moore. Being old doesn't mean acting old: Older users' interaction with spoken dialogue systems. *ACM Transactions on Accessible Computing*, 2(1):1–39, 2009.

[3] K. Georgila, M. Wolters, and J. D. Moore. Simulating the Behaviour of Older versus Younger Users. In *Proceedings of the 46ˆth Annual Meeting of the Association for Computational Linguistics, Human Language Technologies (ACL/HLT)*, pages 49–52, 2008.

[4] K. Georgila, M. Wolters, and J.D. Moore. Learning dialogue strategies from older and younger simulated users. In *Proc. SIGDIAL*, 2010.

[5] R. Vipperla, M. Wolters, K. Georgila, and S. Renals. Speech input from older users in smart environments: Challenges and perspectives. In *Proc. HCI International: Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments*, number 5615 in Lecture Notes in Computer Science. Springer, 2009.

[6] R. Vipperla, S. Renals, and J. Frankel. Ageing voices: The effect of changes in voice parameters on ASR performance. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010.

[7] R. Vipperla. *Automatic Speech Recognition for ageing voices*. PhD thesis, School of Informatics, University of Edinburgh, 2011.

[8] M. Wolters, P. Campbell, C. DePlacido, A. Liddell, and D. Owens. Making Synthetic Speech Accessible to Older People. In *Proceedings of the 6th ISCA Workshop on Speech Synthesis, Bonn, Germany*, pages 288–293, August 2007.

[9] M. Wolters, P. Campbell, C. DePlacido, A. Liddell, and D. Owens. Adapting Speech Synthesis Systems to Users with Age-Related Hearing Loss. In *Beiträge der 8. ITG Fachtagung Sprachkommunikation*, September 2008.

[10] R. Arking. *Biology of Aging: Observations and Principles*. Oxford University Press, 2006.

[11] P. Rabbitt and M. Anderson. Lifespan Cognition: Mechanisms of Change. pages 331–343. Oxford University Press, 2006.

[12] I. J. Deary, M. C. Whiteman, J. M. Starr, L. J. Whalley, and H. C. Fox. 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 (1):130–147, 2004.

[13] J. L. Horn. The theory of fluid and crystallized intelligence in relation to concepts of cognitive psychology and aging in adulthood. In F. I. M. Craik and S. Trehub, editors, *Advances in the Study of Communication and Affect. Aging and Cognitive Processes*, chapter 8, pages 237–278. Plenum Press, New York, NY, 1982.

[14] T. A. Salthouse, R. L. Babcock, and R. J. Shaw. Effects of adult age on structural and operational capacities in working memory. *Psychology of Aging*, 6:118–127, 1991.

[15] P. Verhaeghen. Aging and vocabulary scores: a meta-analysis. *Psychology of aging*, 18:332–339, 2003.

[16] M. A. Shafto, D. M. Burke, E. A. Stamatakis, P. P. Tam, and L. K. Tyler. On the tip-of-the-tongue: Neural correlates of increased word-finding failures in normal aging. *Journal of cognitive neuroscience*, 19:2060–2070, 2007.

[17] S. E. Linville. *Vocal Aging*. Singular Thomson Learning, San Diego, 2001.

[18] F. H. Bess and L. E. Humes. *Audiology: The Fundamentals*. Lippincott Williams & Wilkins,US, 4 edition, 2008.

[19] S Gordon-Salant and P J Fitzgibbons. Sources of age-related recognition difficulty for time-compressed speech. *Journal of Speech, Language, and Hearing Research*, 44:709–719, 2001.

[20] K. S. Helfer. Aging and the binaural advantage in reverberation and noise. *Journal of Speech and Hearing Research*, 35:1394–1401, 1992.

[21] S. Prosser, M. Turrini, and E. Arslan. Effects of different noises on speech discrimination by the elderly. *Acta Oto-Laryngologica Supplement*, 476:136–142, 1990.

[22] C. Smits, S. E. Kramer, and T. Houtgast. Speech reception thresholds in noise and self-reported hearing disability in a general adult population. *Ear and Hearing*, 27:538–549, 2006.

[23] L.-A. Black, C. McMeel, M. McTear, N. Black, R. Harper, and M. Lemon. Implementing autonomy in a diabetes management system. *Journal of Telemedicine and Telecare*, 11 Suppl 1:6–8, 2005.

[24] T. Giorgino, I. Azzini, C. Rognoni, S. Quaglini, M. Stefanelli, R. Gretter, and D. Falavigna. Automated spoken dialogue system for hypertensive patient home management. *International Journal of Medical Informatics*, 74:159–167, 2005.

[25] T. T. Bickmore and D. Mauer. Modalities for building relationships with handheld computer agents. In *CHI '06 extended abstracts on Human factors in computing systems - CHI '06*, page 544, New York, New York, USA, April 2006. ACM Press.

[26] M. Zajicek and Z. Khin Kyaw. The Speech Dialogue Design for a PDA/Web Based Reminder System. In *Proceedings of the 9th IASTED International Conference on Internet and Multimedia Systems and applications, Hawaii*, pages 394–399, 2005.

[27] R. Coulston, E. Klabbers, J. de Villiers, and J.-P. Hosom. Application of Speech Technology in a Home Based Assessment Kiosk for Early Detection of Alzheimer's Disease. In *Interspeech, Antwerp, Belgium*, pages 2420–2423, 2007.

[28] M. Perry, A. Dowdall, L. Lines, and K. Hone. Multimodal and ubiquitous computing systems: Supporting independent-living older users. *IEEE Transactions on Information Technology in Biomedicine*, 8:258–270, 2004.

[29] L. Dulude. Automated telephone answering systems and aging. *Behavior and Information Technology*, 21:171–184, 2002.

[30] M. Zajicek. A methodology for interface design for older adults. In *Enterprise Information Systems Vi*, pages 285–292. 2006.

[31] P. Rabbitt. Mild hearing loss can cause apparent memory failures which increase with age and reduce with IQ. *Acta Oto-Laryngologica Supplement*, 476:167–175, 1990.

[32] J. Raven, J. C. Raven, and J. H. Court. *Manual for Raven's Progressive Matrices and Vocabulary Scales.* Harcourt Assessment, San Antonio, TX, 1998.

[33] N. Unsworth and R. W. Engle. Individual differences in working memory capacity and learning: evidence from the serial reaction time task. *Memory and Cognition*, 33:213–220, 2005.

[34] D. Wechsler. *Manual for the Wechsler Adult Intelligence Scale-Revised*. The Psychological Corporation, New York, 1981.

[35] P. M. Commarford, J. R. Lewis, J. A. Smither, and M. D. Gentzler. A Comparison of Broad Versus Deep Auditory Menu Structures. *Human Factors*, 50(1):77–89, 2008.

[36] K. Georgila, M. Wolters, J. D. Moore, and R. H. Logie. The MATCH corpus: a corpus of older and younger users' interactions with spoken dialogue systems. *Language Resources and Evaluation*, 44:221–261, 2010.

[37] S. Singh, M. Kearns, D. Litman, and M. A. Walker. Reinforcement learning for spoken dialogue systems. In *Advances in Neural Information Processing Systems*, 2000.

[38] O. Lemon and O. Pietquin. Machine Learning for Spoken Dialogue Systems. In *Proceedings of Interspeech, Antwerp, Belgium*, 2007.

[39] J. Schatzmann, K. Weilhammer, M. Stuttle, and S. Young. A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *Knowledge Engineering Review*, 21:97–126, 2006.

[40] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, 2nd edition, 2008.

[41] R. Vipperla, S. Renals, and J. Frankel. Longitudinal study of ASR performance on ageing voices. In *Proc. Interspeech*, pages 2550–2553, Brisbane, Australia, September 2008.

[42] R. Vipperla, S. Renals, and J. Frankel. Augmentation of adaptation data. In *Proc. Interspeech*, pages 530–533, Makuhari, Japan, September 2010.

[43] M. Gales and S. Young. *The Application of Hidden Markov Models in Speech Recognition*, volume 1. Foundations and Trends in Signal Processing, 2007.

[44] J. M. Liss, G. Weismer, and J. C. Rosenbek. Selected acoustic characteristics of speech production in very old males. *Journal of Gerontology*, 45:2, 1989.

[45] G. Karypis. *CLUTO : A Clustering Toolkit*, 2003.

[46] W.I. Hallahan. DECtalk software: Text-to-speech technology and implementation. *Digital Technical Journal*, 7(4):5–19, 1995.

[47] A. S. House, C. E. Williams, M. H. L. Hecker, and K. D. Kryter. Articulation-Testing Methods: Consonantal Differentiation with a Closed-Response Set. *J. Acoust. Soc. Am.*, 37(1), 1965.

[48] L. E. Humes, K. J. Nelson, D. B. Pisoni, and S. E. Lively. Effects of age on serial recall of natural and synthetic speech. *Journal of Speech and Hearing Research*, 36:634–639, 1993.

[49] S. A. Duffy and D. B. Pisoni. Comprehension of Synthetic Speech Produced by Rule: A Review and Iheoretical Interpretation. *Language and Speech*, 35:351–389, 1992.

[50] C. R. Paris, M. H. Thomas, R. D. Gilson, and J. P. Kincaid. Linguistic cues and memory for synthetic and natural speech. *Human Factors*, 42:421–431, 2000.

[51] P. A. Luce, T. C. Feustel, and D. B. Pisoni. Capacity demands in short-term memory for synthetic and nautral speech. *Human Factors*, 25:17–32, 1983.

[52] J. Al-Awar Smither. The processing of synthetic speech by older and younger adults. In *Proceedings of the Human Factors Society 36th Annual Meeting. Innovations for Interactions, 12-16 Oct. 1992*, pages 190–192, Atlanta, GA, USA, 1992. Human Factors Soc.

[53] B. Sutton, J. King, K. Hux, and D. R. Beukelman. Younger and older adults' rate performance when listening to synthetic speech. *Augmentative and Alternative Communication*, 11(3):147–153, 1995.

[54] K. D. R. Drager and J. E. Reichle. Effects of discourse context on the intelligibility of synthesized speech for young adult and older adult listeners: applications for AAC. *Journal of Speech, Language, and Hearing Research*, 44:1052–1057, 2001.

[55] R. W. Roring, F. G. Hines, and N. Charness. Age differences in identifying words in synthetic speech. *Human Factors*, 49:25–31, 2007.

[56] A. Hunt and A.W. Black. Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. In *ICASSP-96*, volume 1, pages 373–376, Atlanta, Georgia, 1996.

[57] J. Yamagishi, T. Nose, H. Zen, Z. Ling, T. Toda, K. Tokuda, S. King, and S. Renals. Robust speaker-

adaptive HMM-based text-to-speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 17(6):1208–1230, 2009.

[58] J. Yamagishi, C. Veaux, S. King, and S. Renals. Voice banking and reconstruction: Speech synthesis technologies for individuals with vocal disabilities. *Acoustical Science and Technology*, 33(1):1–5, 2012.

[59] Maria K. Wolters, Karl B. Isaac, and Steve Renals. Evaluating speech synthesis intelligibility using Amazon Mechanical Turk. In *Proc. 7th Speech Synthesis Workshop (SSW7)*, 2010.

[60] A. P. Breen and P. Jackson. Non-uniform Unit Selection and the Similarity Metric within BT's Laureate TTS System. In *Proceedings of the Third ISCA Speech Synthesis Workshop, Jenolan Caves*, pages 201–206, 1998.

[61] L. Lines and K. S. Hone. Multiple voices, multiple choices: Older adults' evaluation of speech output to support independent living. *Gerontechnology*, 5(2):78–91, 2006.

[62] L. Lines and K. S. Hone. Older Adults' Evaluation and Comprehension of Speech as Domestic Alarm System Output. In *Proceedings of HCI 2002, Human Computer Interaction 2002, Memorable Yet Invisible, London UK.*, pages 94–97, 2002.

[63] B. Langner and A.W. Black. Using Speech In Noise to Improve Understandability for Elderly Listeners. In *Proceedings of ASRU, San Juan, Puerto Rico*, 2005.

[64] N. Unsworth and R.W. Engle. Simple and complex memory spans and their relation to fluid abilities: Evidence from list-length effects. *Journal of Memory and Language*, 54:68–80, 2006.

[65] A. Boothroyd. Developments in speech audiometry. *British Journal of Audiometry*, 2:3–10, 1968.

[66] R.W. Keith. *The Random Gap Detection Test*. St. Louis, 2000.

[67] British Society of Audiology. Pure tone air and bone conduction threshold audiometry with and without masking and determination of uncomfortable loudness levels, 2004.

[68] M. McGee-Lennon, M. Wolters, and S. Brewster. User-Centred Multimodal Reminders for Assistive Living. In *CHI '11: Proceedings of the 29th international conference on Human factors in computing systems*, 2011.