THE EFFECT OF HEARING LOSS ON THE INTELLIGIBILITY OF SYNTHETIC SPEECH

Wolters, Maria[1]*; Campbell, Pauline*[2]*; DePlacido, Christine*[2]*; Liddell, Amy*[1,2]*; Owens, David*[1,2]

(1) Centre for Speech Technology Research, University of Edinburgh
(2) Audiology Division, Queen Margaret University
mwolters@inf.ed.ac.uk; pcampbell|cdeplacido|06006484|06005471@qmu.ac.uk

ABSTRACT

Many factors affect the intelligibility of synthetic speech. One aspect that has been severely neglected in past work is hearing loss. In this study, we investigate whether pure-tone audiometry thresholds across a wide range of frequencies (0.25–20kHz) are correlated with participants' performance on a simple task that involves accurately recalling and processing reminders. Participants' scores correlate not only with thresholds in the frequency ranges commonly associated with speech, but also with extended high-frequency thresholds.

Keywords: ageing, speech synthesis, intelligibility, audiometry, memory

1. INTRODUCTION

Older people are a key user group for a wide range of voice interfaces, including applications such as smart home and home care systems [13], automatic reminder systems [15] and systems for delivering health care interventions [2]. However, hearing abilities decline with age [20]. For example, hearing thresholds, in particular for higher frequencies, increase [9]. Therefore, effects of auditory ageing need to be controlled for in any intelligibility study of older listeners. In this paper, we present the first systematic investigation of the effects of age-related changes in hearing thresholds on the intelligibility of unit selection speech synthesis that covers both conventional pure-tone audiometry and extended high frequencies.

1.1. Influence of Ageing on Intelligibility

Comparatively little work has been done on the relation between hearing difficulties and the intelligibility of synthetic speech, with most of the existing work (e.g. [7]) focussing on formant synthesis and not on concatenative synthesis. Langner and Black [8] asked participants to transcribe sentences which were produced by a human recorded in silence (natural speech), a human who was recorded while listening to a multi-speaker babble (natural speech in noise), a synthetic voice (unit selection synthesis), and the synthetic voice modified to sound like the natural speech in noise. While older listeners (60–90+) performed best in the natural speech in noise and synthetic speech conditions, older participants with self-reported hearing difficulties performed significantly worse than participants with no self-reported hearing problems. Roring, Hines, and Charness [16] report that older subjects performed consistently worse than younger subjects when listening to synthetic speech produced by a diphone synthesiser, even in the presence of context. However, this age effect vanished when hearing loss was taken into account (binaural audiogram, frequencies: 0.25, 0.5, 1, 2, 3, 4, 6, and 8 kHz).

In addition to hearing difficulties, older people tend to have a lower working memory capacity [17], which affects their ability to briefly store information needed for cognitive processing. Even though synthetic speech produced by formant synthesis can be just as intelligible as natural speech on the segmental level [7], it has been shown that this type of synthetic speech can stretch limited cognitive resources even further [10]. This could be due to a number of factors, including lack of acoustic information in the signal [5] and missing or wrong prosodic cues [12]. Auditory stimuli that are difficult to process lead to increased cognitive load, leaving fewer resources for working memory [14, 6]. Although later studies of concatenative speech synthesis have failed to replicate this effect [18, 16], cognitive factors clearly need to be taken into account as potential confounders.

1.2. Hypotheses

In this pilot study, we investigate the effect of hearing loss on participants' ability to understand synthetic speech as produced by a unit selection speech synthesis system, Cerevoice [1]. Although unit selection technology is state-of-the-art in most commercial synthesis systems, there is next to no research on its intelligibility for older listeners, with one main exception [8]. We measured pure-tone hearing thresholds for pure tones with a frequency

range of 0.25–20 kHz, covering both more apical (lower frequencies) and more basal regions (higher frequencies) of the cochlea. We decided to include thresholds for higher frequencies $>$ 8kHz, which are usually not assessed, because the more basal region of the cochlea is often first and most severely affected by the affects of ageing [9]. Moreover, it has been hypothesised that hearing loss in extended high frequencies may be an indicator of early, subclinical damage in more apical regions of the cochlea [11]. To the best of our knowledge, the effect of extended high frequency hearing loss on the intelligibility of synthetic speech has never been investigated before. The hypotheses we are testing here are as follows:
Intelligibility Difference: Synthetic stimuli are less intelligible than natural speech stimuli.
Effect of Hearing Loss: Hearing loss affects the intelligibility of synthetic and natural speech.
Effect of Working Memory: Working memory capacity will explain some of the variation in scores not covered by hearing loss.

2. METHOD

2.1. Participants

This paper reports results from 35 participants recruited for a larger study of auditory ageing. 12 participants were younger (age 25.5 $\pm$ 5 years), 23 participants were aged between 50 and 70 (age 57 $\pm$ 6 years). Starting our older group at age 50 provides a good baseline for subsequent work: Even though there is often already significant hearing loss in the high frequencies in the 50–60 age group, clinical hearing problems are still relatively rare. All participants would pass the pre-experiment screening that is typically used in speech synthesis experiments: They had an average hearing threshold of 20dB or better as averaged over 0.5, 1, 2, and 4 kHz.

2.2. Audiological and Cognitive Tests

In this section, we present only that part of the full assessment battery which is relevant to the results discussed here. All participants completed a working memory test [19] that was presented visually and scored from an answer sheet. Visual presentation was chosen because auditory presentation might affect scores. Pure-tone (PTA) and extended high-frequency (EHF) audiometry was measured on a recently calibrated audiometer (Grason-Stadler, Milford, NH; model GSI 61) in a double-walled soundproofed room (Industrial Acoustics Corporation, Staines, Middlesex, UK). Air-conduction thresholds were measured for each ear at 0.25, 0.5, 1, 2, 3, 4, 6, and 8 kHz following the procedure recommended by the British Society of Audiology [3]. EHF thresholds were established at 9, 10, 11.2, 12.5, 14, 16, 18, and 20 kHz. If a threshold for a frequency f could not be established, we used the maximum intensity for f+5 dB. Testing always began with the better ear in all subjects. Since there are significant differences between the two ears, data from the right and the left ear will be reported separately in this analysis.

For each ear, we computed average hearing thresholds for four frequency groups:
Trad: 0.5, 1, 2, and 4 kHz, the frequencies conventionally used for screening participants in speech synthesis experiments
F1: 0.25, 0.5, and 1 kHz, the frequency range of F1
F2: 1, 2, and 3 kHz, the frequency range of F2
EHF: 9–20 kHz, the complete EHF range.

2.3. Synthesis Experiment

The intelligibility of synthesis systems has traditionally been evaluated using highly artificial stimuli. For this study, we used stimuli that are closely modelled on a real-life application—task reminders. 32 reminders were generated, 16 reminders to meet a person at a given time, and 16 reminders to take medication at a given time. In each group, time preceded person or medication in eight sentences, with the order reversed in the other eight. This allowed us to systematically vary the difficulty of target stimuli, with times being the easiest, person names of medium difficulty, and medication names the most difficult. Person names were monosyllabic CVC words with both consonants being nasal or oral stops, the only exception being the name “Rick”. Names had been chosen to ensure the existence of at least two other names with the same rhyme (-VC sequence). Medication names were constructed using morphemes taken from actual medication names to yield words of 3-4 syllables that did not resemble any existing or commonly used medication. While person names were phonologically simple, but easily confoundable [4], medication names were both unfamiliar and phonologically complex, making them very difficult to remember. This was intended as a safeguard against ceiling effects.

All 32 reminders were synthesised using the Scottish female voice “Heather”. Medication names were transcribed by hand, with the transcriptions adjusted to render them maximally intelligible. The reminders were also read by the same speaker who provided the source material for the synthetic voice. The natural speech was then postprocessed to match the procedures used for creating synthetic speech. The sampling rate of all speech stimuli was 16 kHz.

Participants were asked to recall one aspect of the reminder, either the time or the person/medication.

In each list, participants recalled 16 times, 8 persons, and 8 medications. 8 times, and 4 person and 4 medication names were presented using natural speech, the other 16 were presented using synthetic speech. If participants' responses were a valid pronunciation of the orthographic form of the target word, a score of 1 was assigned, otherwise, a score of 0 was assigned. This procedure takes into account differences in accent between the participants and the Scottish English voice that produced the reminders, such as rhoticity.

Four stimulus lists were created. Each reminder was presented using the synthetic voice in two lists, and using natural speech in the remaining two. Reminders were followed by a short question, recorded using the same natural voice as that used for the reminders. In order to control for recency effects, in two lists (one synthetic, one natural), participants were asked for the first item of a given reminder, while in the other two conditions, participants were asked for the second item. The sequence of reminders was randomised once and then kept constant for all lists.

3. RESULTS

Four scores were computed for each participant: sum of scores for all reminders (**Total**, maximum: 32), sum of scores for reminders presented by a human voice (**Natural**, maximum: 32), sum of scores for reminders presented by the synthetic voice (**Synthetic**, maximum: 16), and difference between the scores for natural and synthetic speech (Δ**(Syn,Nat)**). As Table 1 shows, synthetic speech is more difficult to understand than natural speech ($p<0.5 * 10^{-5}$, V=29, Mann-Whitney test). The main source of the difference are medication names: some medication names are consistently more difficult to understand in the synthetic version than in the natural speech [21]. Although the two groups do not differ in their ability to understand the human reminders, older listeners have significantly more problems with the synthetic voice than younger listeners (two-sample t-test). Could hearing loss be behind this age effect? Even though all participants would have passed traditional initial screening tests, with threshold **Trad** above 20kHz for the better ear, thresholds are significantly higher for the older group ($p<0.0001$ or better, t-test, for all thresholds; cf. Table 2). Table 3 shows raw correlations between the four thresholds and the four target scores for the left ear (*: $p<0.01$, **: $p<0.005$; $p<0.05$ not reported due to large number of correlations computed). We do not report results for the right ear due to lack of space. Looking at the table, we see that the higher **EHF** thresholds and higher **F2** thresholds, the more difficult it is for a listener to understand the synthetic reminders. **F2** covers the frequency range of F2, while **EHF** is computed across the extended high frequencies. In contrast, there are no links between a listener's ability to understand the reminders presented in the human voice and the hearing thresholds we investigated. Judging solely from the spectrum of the speech that was presented, **EHF** should be completely irrelevant, since the highest frequency present in a signal sampled with 16 kHz is 8 kHz. Our hypothesis about the *effects of working memory*, on the other hand, needs to be rejected. Even though there were significant differences in working memory span scores between the two participant groups (younger: 38.9 ± 5.6, 50–70: 27.8 ± 8.6, maximum score: 42), Working Memory Span did not correlate significantly with any of the four scores.

Table 1: Mean scores for each group

Group	Total	Synthetic	Natural	Δ
Younger	28.00 (1.06)	13.50 (1.06)	14.50 (0.53)	-1.00 (1.30)
50–70	26.48 (1.68)	12.30 (1.65)	14.17 (1.31)	-1.87 (2.47)
Sig.	$p<0.05$	$p<0.005$	$p<0.4$	$p<0.1$

Table 2: Average Hearing Thresholds per Frequency Group, left ear, in dB (std. dev.)

Group	Trad	F1	F2	EHF
Younger	0.00 (3.72)	0.17 (2.76)	1.25 (4.44)	9.38 (16.38)
50–60	10.96 (3.93)	5.3 (4.60)	14.20 (5.79)	38.19 (14.20)

Table 3: Correlations between thresholds and audiological measures

Threshold	Trad	F1	F2	EHF
Score	**Left Ear**			
Total	-0.43	-0.22	-0.45*	-0.30
Natural	-0.20	-0.16	-0.15	0.05
Synthetic	-0.46*	-0.19	**-0.53****	**-0.49****
Δ(Syn,Nat)	-0.23	-0.03	-0.32	-0.44*

4. CONCLUSION

To our knowledge, this is the first study to assess the influence of conventional and extended high frequency audiometry on the understanding of

synthetic speech. Our data lead us to conclude that extended high frequency hearing loss ($>$ 8 kHz) clearly predicts the intelligibility of synthetic speech, even for participants who may not normally regarded as "older" (50–60 year-olds) and participants who pass a simple standard screening. EHF thresholds, which are not usually measured, emerge as a strong predictor of participant performance, in addition to a threshold below 5 kHz which covers the regions of the second formant. The correlation between EHF thresholds and participant scores clearly needs to be investigated further. The existence of this correlation and the large amount of unexplained variation both indicate that aspects of hearing loss other than pure-tone thresholds need to be investigated. In future experiments, we plan to compare different types of synthesis systems in order to relate our results to previous work and examine the influence of familiarity and phonological complexity of targets.

5. ACKNOWLEDGEMENTS

This research was funded by the EPSRC/BBSRC initiative SPARC and by the SFC grant MATCH (grant no. HR04016). We would like to thank our participants, our two reviewers for their comments, M. Aylett and C. Pidcock for their invaluable help with generating the stimuli, and R. C. Vipperla for his generous help with digitising minidiscs.

6. REFERENCES

[1] M. A. Aylett, C. J. Pidcock, and M. E. Fraser. The cerevoice blizzard entry 2006: A prototype database unit selection engine. In *Proceedings of Blizzard Challenge Workshop, Pittsburgh, PA*, 2006.

[2] T. Bickmore and T. Giorgino. Health dialog systems for patients and consumers. *J. Biomed. Inform.*, 39(5):556–571, 1006.

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

[4] J. R. Dubno and H. Levitt. Predicting Consonant Confusions from Acoustic Analysis. *Journal of the Acoustical Society of America*, 69:249–261, 1981.

[5] S. Duffy and D. Pisoni. Comprehension of synthetic speech produced by rule: A review and iheoretical interpretation. *Language and Speech*, 35:351–389, 1992.

[6] L. E. Humes. Factors underlying the speech-recognition performance of elderly hearing-aid wearers. *Journal of the Acoustical Society of America*, 112:1112–1132, 2002.

[7] L. E. Humes, K. J. Nelson, and D. B. Pisoni. Recognition of synthetic speech by hearing-impaired elderly listeners. *Journal of Speech and Hearing Research*, 34:1180–1184, 1991.

[8] 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.

[9] F. S. Lee, L. J. Matthews, J. R. Dubno, and J. H. Mills. Longitudinal study of pure-tone thresholds in older persons. *Ear Hear*, 26:1–11, 2005.

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

[11] O. D. Murnane and J. K. Kelly. The effects of high-frequency hearing loss on low-frequency components of the click-evoked otoacoustic emission. *J Am Acad Audiol*, 14:525–33, 2003.

[12] 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.

[13] 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.

[14] M. K. Pichora-Fuller, B. A. Schneider, and M. Daneman. How young and old adults listen to and remember speech in noise. *J.Acoust.Soc.Am.*, 97:593–608, 1995.

[15] J. Pineau, M. Montemerlo, M. Pollack, N. Roy, and S. Thrun. Towards robotic assistants in nursing homes: challenges and results. *Robotics and Autonomous Systems*, 42:271–281, 2003.

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

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

[18] G. Sonntag, T. Portele, and F. Haas. Comparing the comprehensibility of different synthetic voices in a dual task experiment. In *Proc. Third ESCA Workshop on Speech Synthesis, Jenolan Caves*, pages 5–10, 1998.

[19] N. Unsworth and R. 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.

[20] J. F. Willott. *Ageing and the Auditory System*. Singular, San Diego, CA, 1991.

[21] M. Wolters, P. Campbell, C. dePlacido, A. Liddell, and D. Owens. Making synthetic speech more intelligible for older people. submitted.