The Role of Outer Hair Cell Function in the Perception of Synthetic versus Natural Speech

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

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

Abstract

Hearing loss as assessed by pure-tone audiometry (PTA) is significantly correlated with the intelligibility of synthetic speech. However, PTA is a subjective audiological measure that assesses the entire auditory pathway and does not discriminate between the different afferent and efferent contributions. In this paper, we focus on one particular aspect of hearing that has been shown to correlate with hearing loss: outer hair cell (OHC) function. One role of OHCs is to increase sensitivity and frequency selectivity. This function of OHCs can be assessed quickly and objectively through otoacoustic emissions (OAE) testing, which is little known outside the field of audiology. We find that OHC function affects the perception of human speech, but not that of synthetic speech. This has important implications not just for audiological and electrophysiological research, but also for adapting speech synthesis to ageing ears.

Index Terms: speech synthesis, intelligibility, otoacoustic emissions, pure-tone audiometry

1. Introduction

Many factors affect the intelligibility of synthetic speech. One aspect that has been severely neglected in past work is hearing loss. It is well-known that hearing abilities decline with age [1]. This decline can manifest itself in various ways. Notably, hearing thresholds, in particular for higher frequencies, increase [2]. Hearing problems have a significant impact on the intelligibility of synthetic speech. In a study using diphone synthesis, Roring et al. [3] found that a simple pure-tone audiometry hearing threshold explained all age-related intelligibility differences. Langner and Black [4], using unit-selection synthesis, found that older participants with self-reported hearing problems performed significantly worse than those with self-reported normal hearing. We clearly need to investigate how hearing affects the intelligibility of synthetic speech.

Research into this relationship needs to address two issues. First, hearing is a complex process involving sensory components and intricate neural processing. Secondly, most of the data we have about the perception of synthetic speech was generated using formant synthesisers (e.g. [5]) and diphone synthesisers; [4] is an exception. Unit selection speech is spectrally far richer than speech produced by formant synthesis. It has more natural microprosody and far fewer joins than diphone synthesis, and due to the relative lack of signal processing, there are fewer distortions in the speech signal. Hence, we cannot simply transfer results based on one type of synthesis to the other.

In previous work [6], we have shown that the result of [3] also holds for speech produced by unit-selection synthesis: It is not age that predicts inter-individual variation in intelligibility, but hearing thresholds measured using PTA. By "prediction", we mean that selected hearing thresholds explained a statistically significant proportion of the total variance in subject scores. But since PTA effectively probes the complete auditory pathway, it is a blunt tool in the assessment of hearing. Therefore, PTA needs to be supplemented with other assessments that can pinpoint specific problems in the outer ear, middle ear, cochlea, or neural pathway. In this study, we chose to focus on one aspect of the cochlea: outer hair cell (OHC) function. One role of the OHCs is to increase frequency selectivity and sensitivity, acting as a kind of "cochlear amplifier". If damage to the OHCs affects how well participants can understand synthetic speech, then the remedy is relatively straightforward—a filter that approximates the frequency response of this "amplifier", hence compensating for deficits to a certain extent. We are able to examine OHC integrity using an objective measurement: otoacoustic emissions (OAEs, [7]). OAEs are a powerful objective measure of hearing: Not only can they be used to predict type of hearing loss [8], but they may also be sensitive to subclinical damage to the cochlea [9]. In this study, we assess whether OHC function affects how well participants can understand synthetic speech as opposed to natural speech. If it does, then thresholds based on OAE data should cover a significant amount of variation in intelligibility scores.

2. Audiological Background: PTA vs. OAE

2.1. Pure tone audiometry

Pure tone audiometry (PTA) is a psychoacoustic procedure that describes auditory sensitivity. It involves asking a subject to indicate whether they have heard a tone. Hearing threshold levels can be determined by *air conduction* and *bone conduction* audiometry. In air conduction audiometry, the test signal is presented to the test subject by earphones. In bone conduction audiometry, the test signal is presented by a bone vibrator placed on the mastoid or forehead of the test subject. Conventional audiometry tests frequencies between 0.25 and 8 kHz, while high-frequency audiometry tests frequencies between 9 and 20 kHz. Behavioural methods such as PTA can be significantly influenced by a variety of subjective factors including patient co-operation, perception, motivation, attention, linguistic abilities and motor skills as well as subtle cognitive and memory functions. In addition, current reference data suffer from various inadequacies including differing selection criteria, measuring conditions and differing references for normal hearing [10].

2.2. Otoacoustic emissions

Otoacoustic emissions are typically categorised according to the stimulus used to evoke them. In this paper, we will focus on distortion-product otoacoustic emissions (DPOAEs, [11]). DPOAEs are evoked by two pure tones t_1, t_2 with frequencies f_1 and f_2 which are presented at the same time. Due to non-linear processes within the cochlea, responses are created at intermediate frequencies, with a particularly strong one at $2f_1$-f_2. The response at this frequency is called the *distortion product* (DP). The stronger the response, i.e. the higher the DP, the better the amplification. DPOAEs can be tested for frequencies between 1 and 8 kHz and ears with a hearing loss of less than 55 dB as measured by the average PTA threshold for 0.5, 1, 2, and 4 kHz. Audiograms derived from DPOAEs are closely correlated with standard subjective audiograms as measured by PTA [12].

3. Data: Intelligibility of Synthetic Speech

3.1. Cognitive Test

All participants completed a working memory test [13] that was presented visually and scored from an answer sheet. Visual presentation was chosen because auditory presentation might affect scores. Working memory span (WMS) was tested because the experimental task involved remembering the information presented in reminders (cf. Section 3.4 for more detail).

3.2. Audiological Tests

3.2.1. Pure-Tone Audiometry

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 [14]. EHF thresholds were established at 9, 10, 11.2, 12.5, 14, 16, 18, and 20 kHz. 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, 10, and 11.2 kHz, the thresholds measurable in all subjects.

3.2.2. Otoacoustic Emissions

Distortion product otoacoustic emissions (DPOAEs) were recorded for seven f_2 frequencies, 1, 1.5, 2, 3, 4, 6, and 8 kHz, with a f_2:f_1 ratio of 1.22. Both frequencies were presented at 70 dB SPL. This corresponds to distortion product $2f_1$-f_2 frequencies of 0.639, 0.959, 1.279, 1.918, 2.557, 3.836, and 5.114 kHz Responses for these seven DP values were distilled into three variables:
F2DP: The first four DPs, with f_2 covering the range of the second formant
HighDP: The three highest DPs (f_2 4, 6, and 8 kHz), which correlate with extended high-frequency hearing loss [9].
AllDP: All seven DPs (baseline)

3.3. Participants

Two groups of participants were recruited: people aged 20–30 and people aged 50–60. These participants were recruited for a larger study of the impact of auditory ageing on the intelligibility of synthetic speech. In order to control for potential confounders, we eliminated participants with middle ear damage as assessed using tympanometry, which would have invalidated OAE results (n=6), and participants with mild or moderate hearing loss in one or both ears (n=3). As a consequence, any effects of hearing loss on speech intelligibility reported in this paper will reflect the effect of *subclinical* loss. 23 participants remained, 12 younger (2 male, 10 female), 11 between 50 and 60 (3 male, 8 female).

3.4. Synthesis Experiment

For this study, we used stimuli that are closely modelled on a real-life application—task reminders. Task reminders were chosen because this research was partly sponsored by a MATCH (Mobilising Advanced Technology for Care at Home), a multi-centre home care research project, and because task reminders are part of many relevant applications, ranging from electronic diaries to cognitive prosthetics [15].

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. There were three categories of target stimuli, times (easiest), person names (medium difficulty), and medication names (most difficult). Person names were monosyllabic CVC words that had been designed to be easily confoundable [16]. Medication names were constructed using morphemes taken from actual medication names to yield phonologically complex words of 3-4 syllables. This was intended as a safeguard against ceiling effects. Care was taken to ensure that the medication names did not resemble any existing or commonly used medication to avoid familiarity effects.

For the *synthetic speech* condition, all 32 reminders were synthesised using Scottish female voice "Heather" of the unit selection speech synthesis system Cerevoice [17]. Medication names were added to the lexicon before synthesis to eliminate problems due to letter-to-sound rules. The transcriptions were adjusted to render them maximally intelligible. No other aspects of the synthetic speech were adjusted.

For the *natural speech* condition, the reminders were 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. They were first high-pass filtered with a cut-off frequency of 70 kHz, then downsampled to 16kHz, and finally encoded and decoded with the tools `speexenc` and `speexdec`. This procedure eliminates confounding effects due to different voices being used for the synthetic and natural conditions and ensures an exceptionally close matching between human and synthetic speech.

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. Each list was heard by 6 participants, 3 younger (20-30), 3 50–60, except for list D with only two older subjects. The imbalance is due to post-hoc screening.

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. The maximum score possible waas 32; no participant achieved this. This procedure takes into account differences in accent between the participants and the Scottish English voice that produced the reminders, such as rhoticity. Two scores were computed for each participant, each of which was designed to highlight a different aspect of performance:

`Synth:` Σ(scores) for synthesised reminders (processing synthetic reminders)

`Natural:` Σ(scores) for human reminders (processing human reminders)

4. Evaluation: PTA vs. OAE

4.1. Baseline Findings

The two age groups differ significantly along each of the seven thresholds defined in Section 3.2 (Kruskal-Wallis test, $p<0.001$ or better for each ear). Distortion product responses are significantly lower for the 50–60 age group, and PTA thresholds are significantly higher. Table 1 summarises results for four representative thresholds, **TradL**, **TradR**, **AllDPR**, and **AllDPL** (L stands for left ear, R for right ear).

Visual inspection using the R [18] procedure `qqnorm` shows that the four test variables appear to be normally distributed. As Table 2 shows, scores for synthetic speech are slightly lower than those for natural speech. A detailed breakdown of results indicates that this is mainly due to problems with the phonologically complex, unfamiliar medication names.

Before conducting the main tests, we first checked for the effect of potential confounders: list, age group, gender, and WMS (cf. Section 3.1. Age group has no significant effects on any of the scores, even though scores tend to be lower for the 50–60 age group. As we shall see later, though, most of this is due to differences in hearing, as shown by Table 1. There are no significant effects on `Synth`, but both list ($df=3, F=8.56, p<0.001$) and Working Memory Score (WMS, $df=1, F=5.81, p<0.05$) appear to impact on the `Natural` score. The list effect appears to be due to a particularly tricky medication name. The WMS effect is due to two participants with particularly low scores who also had low intelligibility scores.

4.2. Statistical Analysis

In order to assess whether DPOAE data can explain some of the variation in intelligibility scores, we ran four ANOVAs for each target variable, each with a different set of independent variables. We analysed each variable separately because we are interested in comparing the relationship between our independent audiological variables and the intelligibility of human speech to the relationship between our independent variables and the intelligibility of synthetic speech. Our sets of variables are:

PTA-Left: 4 PTA thresholds for left ear

PTA-Right: 4 PTA thresholds for right ear

DPOAE-Left: 3 DPOAE data for left ear, all freqs

DPOAE-Right: 3 DPOAE data for right ear, all freqs

In order to control for the confounders identified earlier, list and WMS were included in all ANOVAS for `Natural`.

Table 3 lists all significant effects that were found ($df=1$, $p<0.05$, except for the list effect on the score `Natural` with $p<0.0001$)). The only factor that explains a significant amount of variation in the score for synthetic speech is **F2L**, the average threshold for 1, 2, and 3 kHz. For `Natural`, however, the overall averages of DP responses for both ears are significant. Since the standard deviation of the audiological variables is much larger compared to the standard deviation of the experimental scores, these correlations might be artefacts of our data set. We are confident that the **F2L** result is solid. Not only has it has been confirmed repeatedly in our companion studies on other subsets of this data set [6, 19], but the F2 frequency range contains much crucial acoustic information in the speech signal. The size of the **AllDP** effect, on the other hand, is much smaller, which means that further validation, in particular through additional experiments, is needed.

In the absence of additional experiments, we checked for artifacts introduced by the data set by constructing linear regression models on three subsets of our data. The subsets were constructed so that each of the 23 participants is omitted from exactly one subset. Each subset contained eight participants in the 20–30 group, and 7–9 participants in the 50–60 group. On each of these subsets, we created four sets of small linear regression models, one per score. For each score, we grew four models by greedy selection. Each model was seeded with one of the four variable pools plus the relevant confounders. Variables were added from the pool until no variable could be found to decrease the model's Akaike Information Criterion (AIC) (procedure `stepAIC`, software package R [18]). The factors that had significant effects on participants' performance according to the main analysis (cf Tab. 3) were almost always included in the relevant models constructed on the three subsets.

Our results indicate a clear difference between natural and synthetic speech: While DPOAE thresholds were related to participants' ability to understand the natural version of the voice, they had no impact at all on participants' ability to understand the synthetic version. Similar findings are evident in the PTA results: the thresholds that were able to predict the intelligibility of synthetic speech are not correlated with the intelligibility of human speech, and vice versa. Since the same speaker was used for both versions of the data, this cannot be accounted for by differences in the speaker herself—it must be due to the different way in which the stimuli were produced. A detailed error analysis [19] suggests that most problems are due to shortened durations. Transitions between the target information and the carrier sentence are particularly crucial.

Table 1: *Age Differences (in dB SPL; mean±stddev))*

Age Group	TradR	TradL	AllDPR	AllDPL
20–30	1.11	0.28	5.00	3.19
	± 6.64	±4.86	±6.09	±5.00
50–60	9.69	11.67	0.26	-0.06
	±4.64	±4.01	±4.84	±5.82

Table 2: *Performance Differences (Age Groups), mean ± stddev*

Score	`Natural`	`Synth`
20–30	14.83 ± 1.11	13.5 ± 1.24
50–60	14.91 ± 1.6	12.73 ± 1.14

Table 3: *Significant Effects*

Score	Pure-Tone Audiometry	
	Left	*Right*
Synth	F2L	none
Natural	List	List,WMS,MidR
Score	**Distortion Product OAE**	
	Left Ear	*Right Ear*
Synth	none	none
Natural	List,WMS,AllDPL	List,WMS,AllDPR

5. Discussion

The two main findings of this paper are: (1) Age-related changes in hearing affect the intelligibility of both human and synthetic speech—even for listeners who would be regarded as "normal" for the purpose of standard speech synthesis experiments. If even sub-clinical changes in hearing affect the intelligibility of synthetic speech, then speech synthesisers must properly adapted to the needs of older ears, or a large number of potential users of synthetic voices will remain excluded. (2) The age-related changes in hearing that affect users' ability to understand synthetic speech may well be different from those affecting the ability to understand natural speech. This finding means that we need to be very careful about extrapolating audiological results about human speech to synthetic speech. Conversely, synthetic speech may be inappropriately used in electrophysiological experiments, resulting in erroneous conclusions. In particular, the condition of the OHCs affects natural speech, but not synthetic speech. However, since the size of that effect is very small, it needs to be validated in further studies. We also plan to consider other relevant aspects of hearing. A promising candidate is central auditory processing, which is affected by auditory ageing [20]. Central auditory processing is involved in compensating for many of the glitches inherent in synthetic speech, ranging from spectral discontinuities to segments which are too short and pitch contours which are unnatural. Finally, we will extend our work to different tasks and participants with a variety of hearing problems.

6. 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 three reviewers for their detailed and insightful 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.

7. References

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

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

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

[4] 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.

[5] 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*, vol. 34, pp. 1180–1184, 1991.

[6] M. Wolters, P. Campbell, C. dePlacido, A. Liddell, and D. Owens, "The effect of hearing loss on the intelligibility of synthetic speech," in *Proc. Int. Congr. Phon. Sci.*, 2007.

[7] D. Kemp, "Stimulated acoustic emissions from within the human auditory system," *J. Acoust. Soc. Am.*, vol. 64, pp. 1386–1391, 1978.

[8] N. Ziavra, I. Kastanioudakis, T. Trikalinos, A. Skevas, and J. Ioannidis, "Diagnosis of sensorineural hearing loss with neural networks versus logistic regression modeling of distortion product otoacoustic emissions," *Audiology and Neuro-Otology*, vol. 9, no. 2, pp. 81–87, 2004.

[9] D. Arnold, B. Lonsbury-Martin, and G. Martin, "High-frequency hearing influences lower-frequency distortion-product otoacoustic emissions," *Arch. Otolaryn. Head Neck Surgery*, vol. 125, no. 2, pp. 215–222, 1999.

[10] M. Buren, B. Solem, and E. Laukli, "Threshold of hearing (0.125-20 khz) in children and youngsters," *Br. J. Audiol.*, vol. 26, pp. 23–31, 1992.

[11] D. Kemp, "Evidence of mechanical nonlinearity and frequency selective wave amplification in the cochlea," *Arch Otorhinolaryngol*, vol. 224, pp. 37–45, 1979.

[12] B. Lonsbury-Martin and G. Martin, "The clinical utility of distortion-product otoacoustic emissions," *Ear and Hearing*, vol. 11, no. 2, pp. 144–55, 1990.

[13] 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*, vol. 54, pp. 68–80, 2006.

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

[15] M. Pollack, "Intelligent Technology for an Aging Population: The Use of AI to Assist Elders with Cognitive Impairment," *AI Magazine*, vol. 26, pp. 9–24, 2005.

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

[17] 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.

[18] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2006, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

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

[20] B. A. Stach, M. L. Spretnjak, and J. Jerger, "The prevalence of central presbyacusis in a clinical population," *J.Am.Acad.Audiol.*, vol. 1, pp. 109–115, 1990.