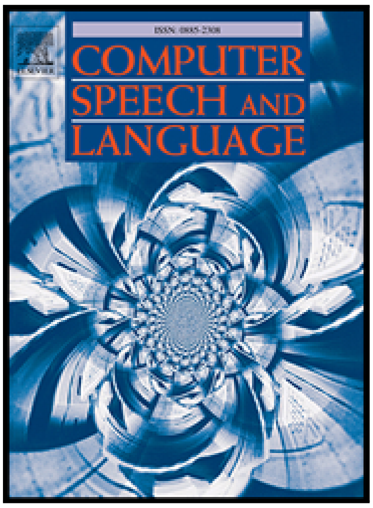

Effects of sentence structure and word complexity on intelligibility in machine-to-human communications

C.T. Justine Hui, Sahil Jain, Catherine I. Watson

PII:	S0885-2308(17)30224-3
DOI:	https://doi.org/10.1016/j.csl.2019.03.002
Reference:	YCSLA 985

To appear in:	*Computer Speech & Language*

Received date:	16 August 2017
Revised date:	18 December 2018
Accepted date:	11 March 2019

Highlights

- Complex syntax and simple words contribute to misunderstandings in listening to unfamiliar synthetic speech
- Created phonetic complexity measure for perception from production based measures
- Effect of how phonetically distinctive the word is is greater than the effect from word occurring frequency in the language on word comprehension
- Significant difference between native and non-native speakers on synthetic speech understanding

Effects of sentence structure and word complexity on intelligibility in machine-to-human communications

C. T. Justine Hui[a,*], Sahil Jain[b], Catherine I. Watson[b]

[a] *Graduate School of Science and Engineering, Sophia University, 7-1 Kioi-cho, Chiyoda-ku, Tokyo 102-8554, Japan*
[b] *Department of Electrical and Computer Engineering, University of Auckland, Auckland 1010, New Zealand*

Abstract

With the rise of robotics and artificial intelligence, good communication between humans and machines becomes more important. However, users with language and hearing disadvantages may find synthetic speech systems to be difficult to understand. In this study, we explore the types of sentence structure and level of word complexity that affect intelligibility of speech in unfamiliar context. Using semantically unpredictable sentences, we found that sentence with more complex syntax such as relative pronouns and question words are harder to comprehend, while on the word level, it is the shorter and simpler words that contribute to misunderstandings. We found that although word frequency affects how well a word is recognised, the effect from the occurring frequency is much less than the effect of how phonetically distinctive the word is. There was also evidence of significant difference between native speakers and non-native speakers on how well they could understand the sentences. These results may help us in designing better dialogue system for machine to human interactions, especially in the healthcare arena, where often users have disadvantages in language and hearing abilities.

Keywords: dialogue design, speech perception, speech synthesis, machine to human communications

1. Introduction

Dialogue design and intelligibility in synthetic speech have long been a subject of interest for human to machine interaction with the rise in robotic study. Studies have mainly concentrated on how to make dialogue spoken by systems more natural and human-like, mimicking human-to-human dialogue in human-to-robot communication (see e.g. Fong et al., 2001; Bartneck and Forlizzi, 2004; Edlund et al., 2008).

[*] Corresponding author; email:justinehui@eagle.sophia.ac.jp

The studies focus on understanding humans and responding in a more human-like way, with the assumption that users can understand the robots without much problem. However, it has also been reported that hearing and language abilities can affect this other side of the interaction, where listeners have trouble interpreting the synthetic speech. Wolters et al. (2007a,b) conducted perception tests on how intelligible synthetic voice in the healthcare context is for hearing impaired participants, and found that elevated auditory threshold greatly affected their understanding of the voices, especially with unfamiliar words and context. Similarly, Igic et al. (2010) have investigated the perception of robotic speech and the importance of familiar accent in the synthetic voice in a healthcare environment. In another study, Watson et al. (2013) found that non-native language status of the listeners greatly affected their understanding of synthetic speech, even when no difference can be seen between the groups for natural speech. These studies suggest the importance of designing synthetic speech for those with disadvantage in listening and second language abilities.

In this study, we explore how different types of sentence structure and word complexity can affect the intelligibility of speech in a semantically unpredictable environment. This may provide guidance for future dialogue design, especially in the healthcare industry where there may be a large percentage of users having the disadvantage from age, hearing loss and second language. While we know that the difficulties in understanding speech from hearing loss and age differ greatly from being a non-native listener, we believe that by increasing the accessibility of dialogue in human-to-machine interactions, we can address these intelligibility related disadvantages and reach out to a wider audience. The current study focuses on the effect of nativity, namely New Zealand English in this case, with future plans to extend the research into age and hearing loss.

2. Methodology

The current paper examines the results from an intelligibility perception test carried out as part of a bigger study to evaluate newly developed synthetic voices in New Zealand English (NZE) (Jain, 2015). The test was conducted over an online platform which involves participants listening to semantically unpredictable sentences (SUS) and typing in what they heard. A web-based approach was chosen to reach out to more participants and the uncontrolled sound environment was deemed acceptable as the synthetic voices are purposed for robots' speakers, where sound quality may be restricted. This is also the default evaluation method carried out for the Blizzard challenges, an annual international competition for building synthetic voices (Fraser and King, 2007). Eight synthetic voices created with recordings from four NZE speakers were used as the stimuli of the sentences. These voices were created in the same way, modelled using Hidden Markov Models on OpenMARY, a text-to-speech (TTS) engine (Schröder and Trouvain, 2003). The two synthetic voices from each speaker differ in training parameters and no significant difference was found between them, and therefore we evaluated the results on a speaker basis instead of each voice. In this paper while the effect of speaker will be taken into account, we will focus the effect from sentence and word complexity.

2.1. Participants

Sixty-two participants carried out the test, with twenty-one female participants and forty-one male participants, age eighteen to seventy-five. The majority of the participants are younger than 45 (95%) and based in New Zealand. Forty participants are identified as native New Zealand English (NZE) speakers (65%) and rest were not native speakers of NZE. We determine a participant to be a native NZE speaker if they were either born in New Zealand, or have moved to New Zealand before the age of ten.

Nativity for the current study is defined as being native to New Zealand English (NZE). New Zealand English is the variety of English spoken by people of European descent in New Zealand Bauer et al. (2007). In general, the phonetic system of New Zealand English is similar to other non-rhotic standard varieties of English, but it also has distinctive features from other English's, such as the centralised New Zealand /ɪ/ vowel (as in "kit") to be the same phoneme as the /ə/ vowel (as in second syllable of "comma"), and a ɪə/eə (as in "near/square") merger for the younger generation.

While other research have focused on non-native speaker being a second language learner compared to native speakers of English, in the current study, we focused on native speakers of one variety of English as all our participants are considered to have a fair command of English. In addition, previous literature have found foreign accent to affect both young and elderly listeners' perception of speech, and in particular, elderly listeners having trouble adapting to foreign accent (Adank and Janse, 2010; Ferguson et al., 2010). With our motivation the healthcare industry, we considered it crucial to investigate from a English variety point of view as both the carers and patients may not be adapted to NZE. Despite varying English experiences and native language in our non-native group, all participants are familiar with English, but not NZE. While we did not have records of any official English assessment scores for language proficiency, 88% of the participants considered themselves to have above-average to excellent command of English, and the rest were either born in New Zealand (and thus were most likely being humble) or university students in New Zealand, and therefore would have passed the English prerequisite for university entrance.

2.2. Semantically Unpredictable Sentences (SUS)

The semantically unpredictable sentence (SUS) test described in the current paper was inspired by the Blizzard Challenge (Fraser and King, 2007; King, 2014; King and Karaiskos, 2016), and was constructed based on the specifications in Benoît et al. (1996). Being in its fourteenth year at the time of writing, the Blizzard challenge is a well established international speech synthesis competition to systematically compare different text-to-speech systems. The SUS test is used in the Blizzard challenges to examine the intelligibility part of the system, that is, how well the synthetic voices can be understood. The sentences are created to be meaningless and hard to comprehend in order to control the semantic information and contextual cues in the stimuli (Benoît et al., 1996; King, 2014). The idea behind using nonsense sentences, according to the Blizzard organisers, is that if the synthesiser is as intelligible as natural speech when using difficult, meaningless sentences then the synthesiser should be at least as intelligible using normal sentences in less than optimal environments and comprehension abilities (Fraser and King, 2007; King, 2014).

The sentences being tested were generated in the format shown by the sample sentences in Table 1. Benoît et al. (1996) developed these 5 sentence types, intending for the sentences to be simple and applicable to other European languages such as Dutch and German. While the Blizzard challenge took their SUS test from Benoît et al. (1996), they used a simpler sentence type in the form of det-adj-noun-verb-det-adj-noun. Other studies have also used SUS to evaluate intelligibility for applications other than synthetic speech evaluation, e.g. cue enhancement in noisy environment (Hazan and Simpson, 1998). However, there does not seem to be any literature comparing all five sentence types.

There were 21 sentence variations overall, and each sentence type has 3 to 5 variations. The participants were given eight sentences randomly, and by the end of the trial, the number of each sentence type tested were 107 for Type 1, 84 for Type 2, 100 for Type 3, 91 for Type 4 and 80 for Type 5. In total, we obtained 462 trials, with an average of 22 trials per sentence variation. Each participant listened to one sentence at a time and typed in what they heard into the web survey. The participants were told prior to the experiment that the sentences may not make sense. The words that made up the sentences were randomly chosen from the top 5000 words in the Corpus of Contemporary American English (Davies, 2010), making sure they were also common words in NZE, for example using "footpath" instead of "pavement".

While the SUS test has been shown to be a good indicator of intelligibility compared with other kinds of intelligibility tests (Benoît et al., 1996), a number of participants in our study expressed concern at only being able to listen to the SUS once, and forgetting the sentence by the time they had finished listening. This suggests that the complexity of the SUS impacted on the intelligibility. In the following sections we investigate intelligibility by describing two factors for analysing the results obtained on these sentences: sentence complexity and word complexity.

3. Sentence complexity

Sentence complexity is one of the more common ways to evaluate the intelligibility of a synthetic voice, analysed using the word error rate as shown in the Blizzard challenges (King, 2014) and other speech recognition evaluations (Benoît et al., 1996) .

3.1. Results

Each sentence was analysed for their word error rate, and three sources of errors were identified:

Table 1: Structure of SUS used in the survey, with an example of each sentence type given

Type	**Word 1**	**Word 2**	**Word 3**	**Word 4**	**Word 5**
1	Noun	Intr Verb	Preposition	Adj	Noun
	The meaning	*travelled*	*with*	*the strong*	*solution*
2	Adj	Noun	Tran Verb	Noun	
	An important	*location*	*fought*	*an investment*	
3	Tran Verb	Noun	Conjunction	Noun	
	Clean	*a story*	*or*	*the payment*	
4	Question	Noun	Tran Verb	Adj	Noun
	Where does	*the secretary*	*turn*	*an interesting*	*television*
5	Noun	Tran Verb	Noun	Rel Pronoun	Intr Verb
	A message	*removed*	*a paper*	*who*	*waited*

- Substitutions - words substituted for an incorrect word
- Deletions - words not recorded
- Insertions - words incorrectly inserted

The following word error rate (WER) metric was used to quantify the proportion of words incorrectly recorded by participants:

$$WER = \frac{S + D + I}{N} \tag{1}$$

Where WER is the word error rate, S the number of substitutions, D the number of deletions, I the number of insertions, and N the total number of expected words in the SUS. A WER was calculated for each sentence the participant listenered to according to Equation 1, resulting is 462 WER scores to evaluate.

In order to look at sentence complexity, we must first define what we mean by the term *complexity.* We examine two sources of sentence complexity in this paper, the sentence type according to how the sentence is structured grammatically, and the number of words in the sentence.

3.1.1. Sentence Type

As described in Table 1, each sentence type differed in their grammatical structure. Figure 1 shows the results for each sentence type and we can see that there is a strong implication that Type 1 sentences are more intelligible than the other sentence types.

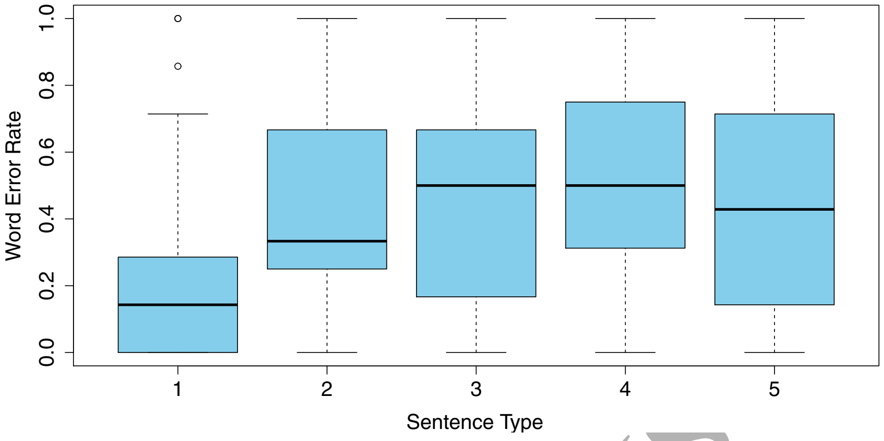

Figure 1: Word error rates relative to sentence type

We used R (R Core Team, 2015) and *lme4* (Bates et al., 2015) to perform a linear mixed effects analysis of the relationship between WER and sentence type. As fixed effects, we entered the sentence type and the speaker of the voice into the model. Visual inspection of the residual plots did not reveal any obvious deviations from homoscedasticity or normality, satisfying assumptions. Evidence of significant differences ($\chi^2(4) = 111.71$, p<0.001) was obtained by a likelihood ratio test. A Tukey multiple comparisons of means test showed significant differences between sentence type 1 and all other sentence types (z_{1-2}=6.07; z_{1-3}=8.81; z_{1-4}=9.8; z_{1-5}=8.77, p<0.001).

Benoît et al. (1996) regarded all SUS types to be simple sentence structures easy enough for non-natives to comprehend. Referring to Table 1, we can see that Type 1 sentences employ a "subject-verb-object" formula, which is more straightforward syntactically compared to Type 3, 4 and 5, where the sentences start with a verb, a question word, or consist of a relative pronoun. It is interesting to note that Type 2 also follows the "subject-verb-object" formula. While there is no evidence of significant differences between Type 2 and the other sentence types, we can observe a lower median compared to the other sentences from Figure 1 .

3.1.2. Words in Sentence

We saw in the previous section that the sentence type affects how well the participants could hear the sentences. As each sentence variation has different number of words, this may also contribute to how well participants performed, where we would expect a longer sentence to be more difficult. Figure 2 shows the result for the number of words in a sentence. This shows that overall, participants found sentences with six or eight words the most difficult to understand, with a median WER of around 0.5. We performed a linear mixed effects analysis for the relationship between WER and number of words in a sentence. As fixed effects, we entered the number of words and speaker of the voice into the model. As the random effect, we had intercepts for the participants. Evidence of significant difference ($\chi^2(2) = 37.79$, p<0.001) was obtained by likelihood ratio tests comparing the full model with both fixed effects against a model with the speaker fixed effect only. A multiple comparisons of Tukey contrasts showed significant differences between 6 and 7 words (z=4.4, p<0.001), between 7 and 8 (z=5.9, p<0.001), but not between 6 and 8 words (z=2.3, p=.055). Therefore, the number of words in the sentence cannot be used to explain how well the participant can understand the stimuli, as opposed to the sentence structure.

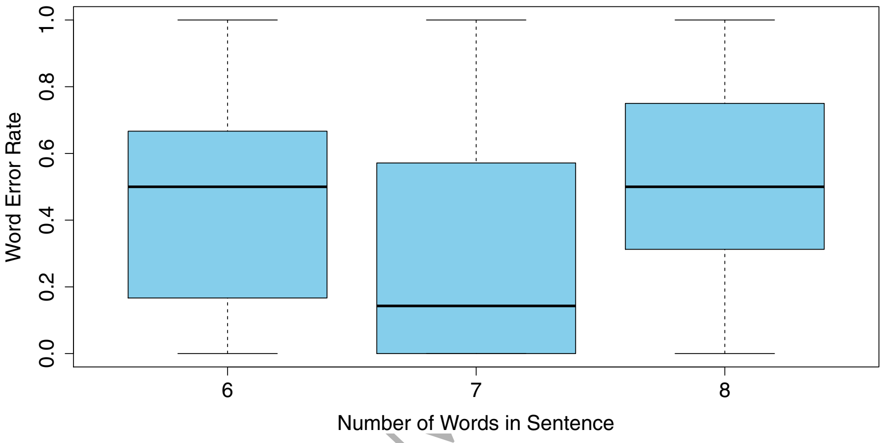

Figure 2: Word error rates relative to number of Words

3.2. Discussion

From Table 2 we can see that the results for seven words in a sentence was affected partially by Type 1 sentences. This suggests that sentence structure affects intelligibility more than the number of words; the simpler the sentence is, the easier it is to perceive in an unfamiliar environment.

Table 2: Number of words for each sentence type

Type	1	2	3	4	5
Words	7	6	6	8	7

While we concluded that simple grammar makes the sentences more intelligible, there was evidence of significant difference between Type 1 and Type 2 sentences, which share a similar grammatical structure. Taking a look at the errors made by the listeners and the stimulus sentences suggest that the transitive verb with a subject and object (e.g. "a pregnant combination terrified an advice ") seem to be more difficult to comprehend than the sentence with an intransitive verb followed by a verb-less clause (e.g. "the meaning travelled with the strong solution"). This could be partly because the sentence can be separated into two shorter clauses, which may be easier to identify despite its meaninglessness, and therefore easier to capture in memory. We did however try to prevent memory saturation by limiting the sentences to 6 to 8 words (Benoît et al., 1996).

Wolters et al. (2007b) found that the position of the words affect how well the participant can comprehend the words and certain positions are easier to understand than others, especially via synthetic voices. While the sentence structure in Wolters' study was not the same as our study, this supports the fact that especially for listening to synthetic voices, there is a heavier load from processing grammar, and the structure of the sentences affect intelligibility.

4. On Word and Phonetic Complexity

The word error rate allowed us to examine the intelligibility of the sentences on a whole. However, we are also interested in how well participants were able to hear each word in the sentences, and this may give us an idea of the type of words that can be easily recognised. To examine word complexity effect on intelligibility, we used word correct rate (WCR) to analyse the results. WCR is simply the number of words the participants correctly answered compared to the total number of words. A higher WCR indicates that there are more participants understanding the words. There were two factors we explored, the number of syllables in the word and how phonetically complex the word is. We created our own measure for this using a similar measure to the word complexity measure (WCM) as developed by Stoel-Gammon (2010) and Jakielski (1998)'s index of Phonetic Complexity (IPC). In total, we had 85 number of unique words in our analysis excluding articles, 2087 responses from the participants, and on average, 25 number of responses per word.

4.1. Results

4.1.1. Number of Syllables in Word

Previous studies have defined phonetic complexity using bisyllabic and monosyllabic words (Lass et al., 1979; Francis and Nusbaum, 1999), with the assumption that the more syllables a word consists of, the more complex phonetically it is. Following this idea, we then analysed the results according to the number of each syllables, excluding articles 'a', 'an' and 'the', and conjunctions 'and' and 'or'. The results can be seen in Figure 3. We can see a clear trend where the more syllables in a word, the higher the accuracy rate.

Figure 4 shows the actual words the participants heard according to the correct rate relative to the number of syllables. (Articles and conjunctions are included in the figures for completeness.) We can observe that some of the monosyllabic words with higher correction rate, such as "strong", are more 'phonetically complex' than 'burn' due to the consonant cluster at the start of

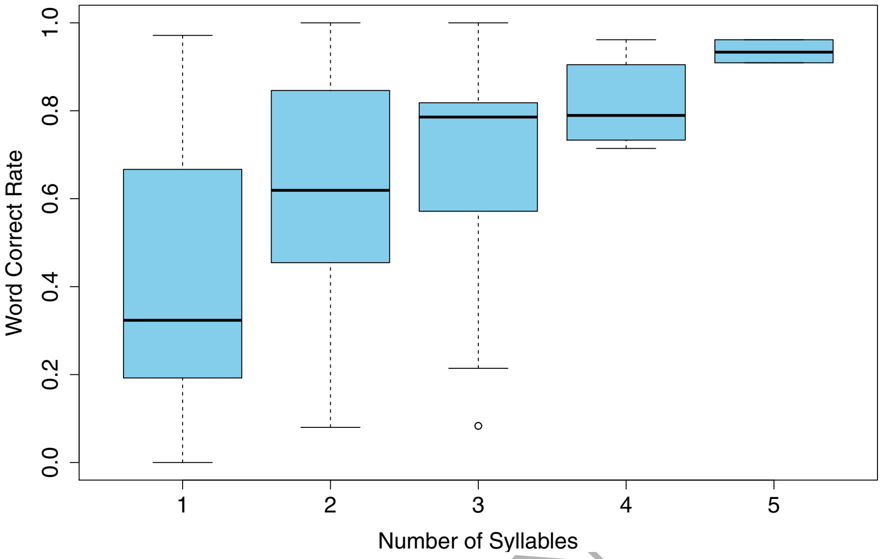

Figure 3: Correct word rate relative to number of syllables

the word. This implies that we need a more phonetically motivated measure to analyse the results.

Analysing with a linear mixed effect model, we have the number of syllables and the speaker of the voice as the fixed effects to observe the results from the binomial data of participant answering correctly or not. Evidence of significant differences ($\chi^2(4)$ = 239.89, p<0.001) was obtained between the number of syllables. Employing a Tukey test showed evidence of significant difference (z_{1-2} = 5.15; z_{1-3}= 9.28; z_{1-4} = 10.78; z_{1-5} = 5.55;z_{2-3} = 4.36; z_{2-4} = 6.94; z_{2-5} = 4.44; z_{3-4} = 3.29;z_{3-5} = 3.22, p<0.01) between all numbers of syllables apart from the penultimate and ultimate number of syllables investigated. This suggests that the more syllables a word has, the easier and less ambiguous it is for participant to identify in non familiar context.

4.1.2. Word Complexity Measure (WCM)

We realised from our above investigation that the number of syllables in the word as a measure of word complexity does not account for phonetic information of the words. At the same time, we can not analyse the words based solely on its phonemes because while the sentences are semantically unpredictable on a whole, each word stands on its own semantically. For example, if a word such as 'pregnant' results in a high correct rate and can be recognised easily, it does not necessarily mean that the consonant clusters 'pr', 'gn' and 'nt' are easier to recognise, but instead, it is the holistic effect of the word as a whole. On the other hand, when we listen to the word 'pregnant', while being only 2 syllables, we can see that the word holds more information phonetically than another 2

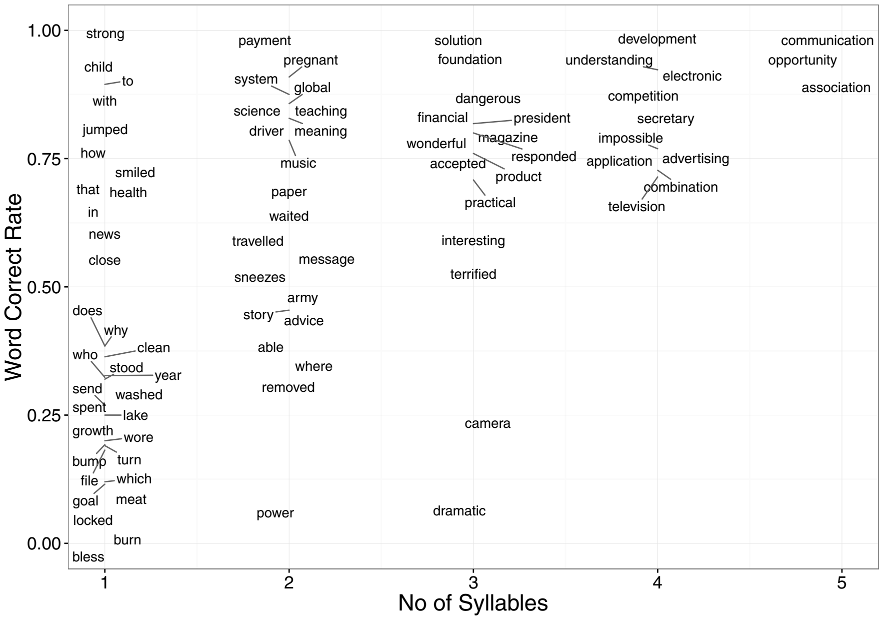

Figure 4: Correct rates of text relative to number of syllables

syllable word such as 'paper'. The phonetic complexity of the word 'pregnant' came mostly from its clusters and the consonant ending. This brings us to investigate for a new set of measures in addition to the number of syllables per word.

Phonetic complexity has been used widely to evaluate the development of speech in children (Chen and Liu, 2015, e.g.), as well as speech disorders such as aphasia (Haley and Jacks, 2014, e.g.), stuttering (Coalson and Byrd, 2016, e.g.) and dyslexia (Bose et al., 2011, e.g.). The two more widely known measures are Jakielski (1998)'s index of Phonetic Complexity (IPC) and Stoel-Gammon (2010)'s word complexity measure (WCM), which both describe and evaluate early acquisition patterns in children's speech.

The criteria in both measures look at the consonant by place and manner, for example fricatives being harder to pronounce than stops and dorsals being harder to say than labials, as well as voicing differences, word shape (whether or not the word ended in a vowel or consonant), number of syllables and consonant clusters. For the purpose of our study, we took these measures which were motivated by speech production, and used some of the criteria for perception evaluation.

The following three points were chosen as the other criteria were deemed not relevant for perception. We did not consider the manner or place of articulation or voicing to affect hearing as much as speaking, and therefore those criteria were not used to evaluate the words. The number of syllables adds to the complexity of the words, as the more syllables the word has, the more phonemes it would contain, and thus more phonetic information in the word. Having a word-final consonant, while being more difficult to pronounce, also adds information to the word, and more difficult to capture as in general it contains less energy than a vowel ending. Lastly, a cluster combines two or more consonants, holding more phonetic information than a single consonant.

- 1 point for each syllable
- 1 point for word-final consonant
- 1 point for each consonant cluster

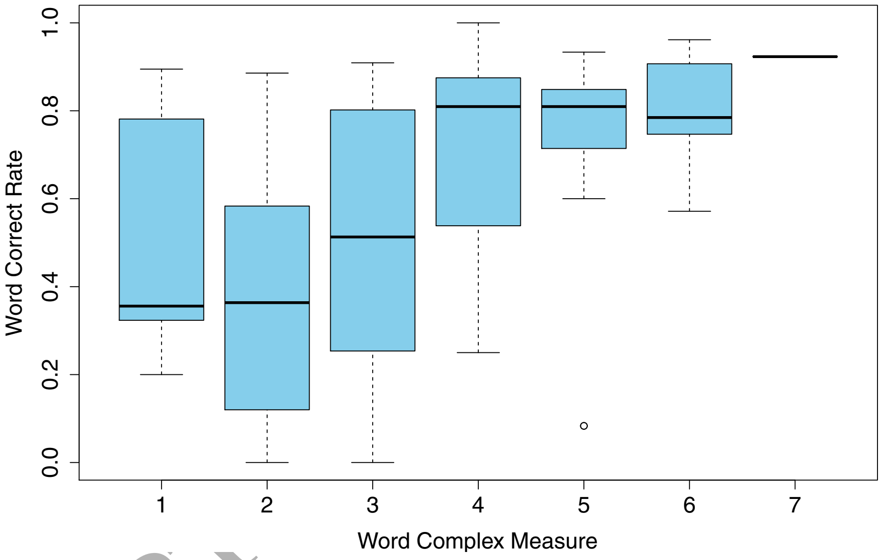

Figure 5: Correct rates for Word Complex Measure

For example, the word ’strong’ would give a WCM of 3, calculated by adding 1 point for one syllable, 1 point for consonant cluster, and one point for ending in a word-final consonant.

We can observe here that the more complex phonetically the word is, the easier it is for listeners to perceive correctly in Figure 5. From Figure 6, we can see the remaining words in the lower WCM which scored relatively high are prepositions, relative pronouns and question words. This may be an effect of occurrence frequency in daily conversation. Analysing with a linear mixed effect model, with the fixed effects as the WCM and the speaker of the voice, we found significant differences ($\chi^2(6) = 274.4$, p<0.001) between the groups of words in terms of WCM. A Tukey analysis gave evidence of significant differences (p<0.05) between all WCM apart from 1 and 3, 4 and 5, 4 and 6, 5 and 6, and between 6 and 7. This shows that while participants find words easier to understand when they are phonetically more complex, the complexity of a word does not increase intelligibility after a certain point, that is, between WCM of 3 and WCM of 4.

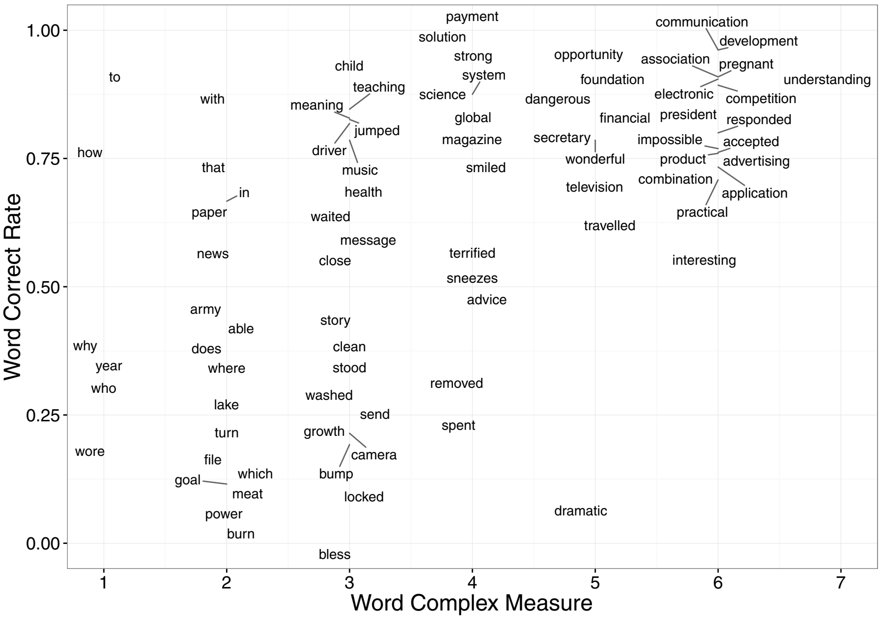

Figure 6: Correct rates of text for Word Complex Measure

4.2. Discussion

Simple words with ambiguous and similar phonetic structures were perceived incorrectly more often. For example, comparing 'pregnant' with 'burn', we have a word where /pr/ is followed by /gn/, a relatively unique combination in English, while 'burn' has many minimal pairs. Words that have a high WCM but low accuracy also resemble parts of a minimal pair, for instance 'dramatic' was heard as 'grammatical'. This shows that while the word needs to be relatively phonetically complex, they also need to be unambiguous to be recognised with ease. The results are not surprising, as phonetically more complex and unambiguous words tend to hold more information than less complex words. As part of analysing the data, we adapted a production focused metric to be used for perception. There does not seem to be any systematic approach to date to look at how complex a word is phonetically from a perception point of view. While most of the literature focused on production and language development, we were able to apply a similar measure to analyse words in terms of its phonetically complexity for listening. Some of the criteria used in Stoel-Gammon (2010)'s

WCM include very specific production properties, such as whether a word includes a velar final stop. These criteria were deemed less important in terms of perception. However, this raises the question of listening to sounds the listeners have difficulty producing, and whether that will make the phonetically complex word easier or harder to capture. A future study can look at how non-native listeners respond to words with phones they are not familiar with, for example, Japanese participants listening to words with /l/ and /r/ such as 'library' and see how that compares with another similar phonetically complex word such as 'magazine'.

While WCM can give a more phonetically based metric than number of syllables for words such as "pregnant", we found number of syllables to predict similarly to WCM in the current study as seen in Figure 3 and Figure 5. Previous literature have found listening in adverse environment such as noise and masked speech to be affected by nativity of the listeners (e.g. Hioka et al., 2017). Hence, more adverse listening conditions may give advantages to using WCM as a metric to determine whether a word could be easily perceived but this was not observable in our results. Further investigations will be needed to examine the effectiveness of using WCM to predict intelligibility.

5. Effect of the Synthetic Voices

Voices from four speakers were used as stimuli for the perception test. While there was no significant difference in terms of the training parameters between the two voices from the same speaker, we saw evidence of speaker effect on how well the participants could recognise the stimuli ($\chi^2(7) = 71.312$, p<0.001). Using a linear mixed effect model analysis, we found that there was no evidence of interaction between sentence type with the different speakers ($\chi^2(14) = 15.6$, p=.34). This means that we were able to interpret the results from sentence complexity independently of the difference in synthetic voices. We also found that the speaker effect (F(3)=10.64)) was lower comparing to the effect from the sentence type (F(4)=47.11). This means that while two of voices we used had higher intelligibility scores than the other two, the sentence effect was considerably stronger.

On the other hand, we observed interaction between the speaker of the voice with both the number of syllables ($\chi^2(11) = 37.37$, p<0.001) and WCM ($\chi^2(24) = 90.97$, p<0.001). This suggests that we may need to reconsider our interpretation of the results regarding word complexity, although the F value for the interaction ($F_{syl}(12) = 2.87$;$F_{wcm}(24) = 3.25$) is also relatively lower than number of syllables (F(4) = 41.47), WCM (F(6) = 33.7) and speaker difference ($F_{syl}(5) = 12.46$; $F_{wcm}(5) = 8.37$). As future work, we should concentrate on one synthetic voice and examine the effect of sentence and word complexity without the effect of different voices.

6. Effect of English Nativity

Watson et al. (2013) suggests that native NZE speakers are more attuned to synthetic NZE speech than non-native NZE speakers. In this section, we compare between the native NZE speaker group with the non-native NZE speakers in terms of their results with sentence and word complexity.

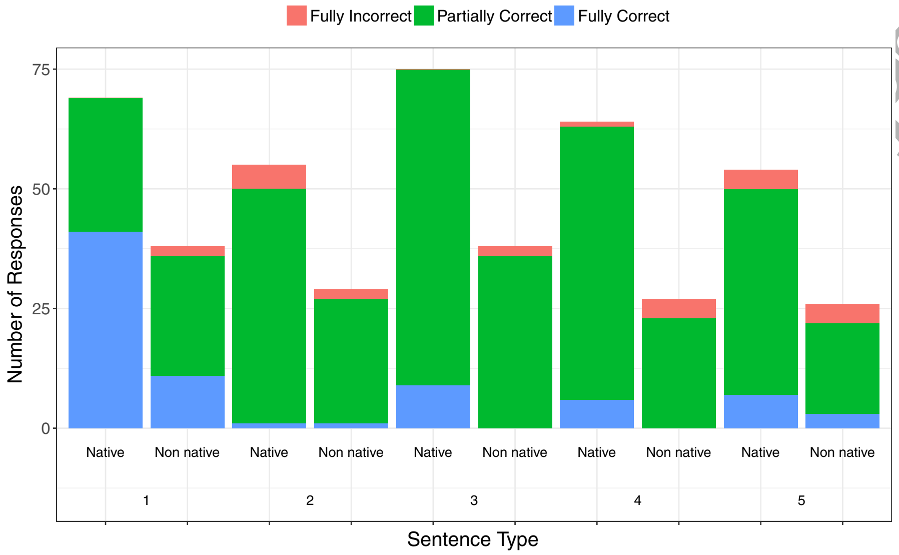

Figure 7: Distribution of responses relative to sentence type for native and non-native speakers

Figure 7 shows the distribution of how well native NZE speakers and non-native NZE speakers could understand the sentences. Native participants were more likely to get responses fully correct, especially for Type 1 sentences. Non-native NZE speakers had more trouble with Type 4 sentences, which starts with a question word. We also see native speakers comprehending individual words slightly better than non-native speakers for the lower number of syllables in Figure 8 and Figure 9.

While there is significant difference between nativity for sentence complexity ($\chi^2(1) = 5.5$, p=.02) and for word complexity ($\chi^2(1) = 6.47$, p=0.01), there is no evidence of interaction between nativity and sentence type, number of syllables and WCM.

This supports the conclusion from Watson et al. (2013), where while there was no difference between L1 and L2 speakers in understanding natural speech, they found a difference when the participants listened to synthetic speech, suggesting that listening to synthetic speech adds extra cognitive load to language processing for L2 speakers. Past Blizzard challenges have also reported on how the SUS part of their tests was favoured by native speakers the most, where they found the nonsense sentences to be amusing. On the other hand, non-natives found it the most difficult section (Fraser and King, 2007).

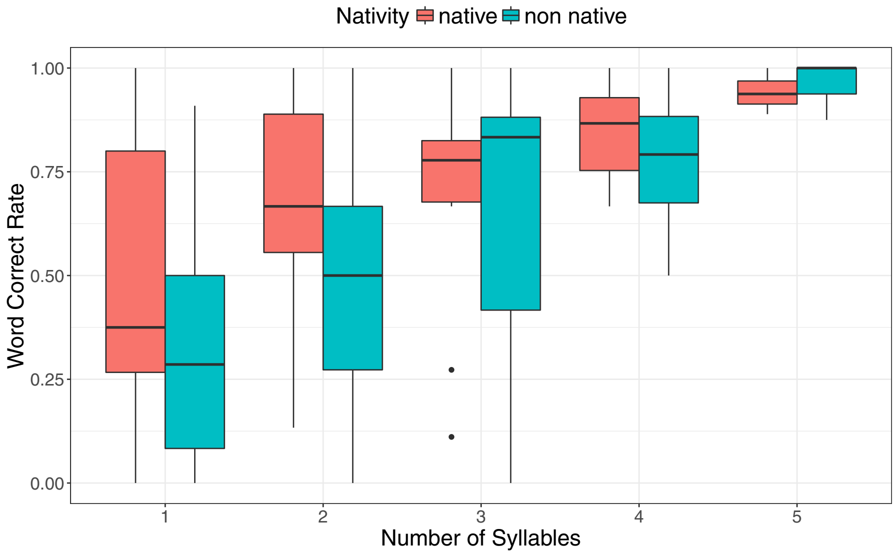

Figure 8: Comparing response correctness of native and non-native speakers relative to word complex measure

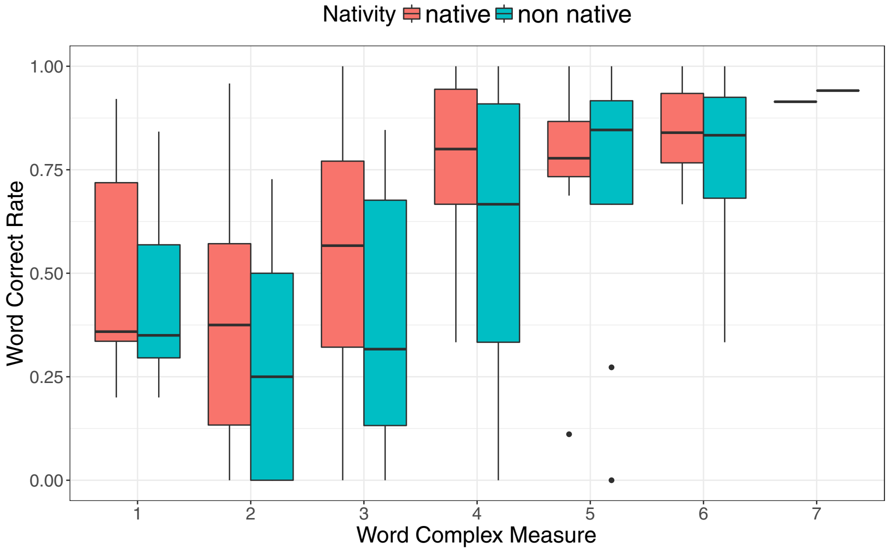

Figure 9: Comparing response correctness of native and non-native speakers relative to word complex measure

7. Effect of Word Frequency

While the sentences were designed to be semantically unpredictable to avoid any cognitive effect, the frequency of the words themselves may make some words easier to understand than others. The list of words were put through the Sublexus database (Brysbaert and New, 2009) and grouped according to their occurring frequency count to examine word frequency effect on their correct rate. Using a linear mixed effect model with the frequency groups as fixed effect, a likelihood ratio test showed that there is significant difference of word frequency on the correct rate ($\chi^2(1) = 19.81$, p<0.001). We then examined the effect of frequency groups with either number of syllables or WCM, by adding into the model either of the factors as fixed effect. Inputting the model into *anova()* from the *lme4* package (Bates et al., 2015) in R however, showed a much greater F value for number of syllables (F(1)=167.36), than for the word frequency groups (F(1)=3.9) (and when comparing between WCM (F(1)=121) and word frequency groups (F(1)=4.37)). We can observe this from the results where the word 'pregnant', a relatively rarer word, was easier to perceive than 'burn', a relatively more common word. This suggests that while word frequency has an effect on how easy it is for the participants to recognise, the effect is much smaller than that from the number of syllables or WCM.

It is also interesting to note that from the current study, while frequency of the word does contribute to how easy the word is to be recognised, it is the phonetic complexity of the words that have a greater effect. This suggests that just because a word is frequently used, does not mean that it can be easily recognised in an unfamiliar environment. However, more testing with a bigger word dataset is required to verify this. On the other hand, we also observed words with affricates tend to be recognised fairly easily (e.g. 'jumped') in the current dataset. This could be attributed from the rarity of affricates in English, but again, we have a very limited set of words in the current study to make an observation. In the previous section we discussed how it may be useful to include the criteria of consonants from Stoel-Gammon and Jakielski's measures for L2 speakers.

Similarly, looking at the consonant sound and their occurring frequency may also give us different insights as the less frequent the sound occurs in the language, the more unique a word with the particular sound would be. Also, the fact that we found lesser frequency effect may be due the current study using the words from the top 5000 common word list. Nevertheless, due to the increased cognitive load for processing synthetic speech, we recommend using familiar words to design the dialogue for synthetic speech systems.

We have mentioned in the beginning of the paper that hearing loss and age also contribute to difficulty in understanding synthetic speech. There are obvious differences in the types of difficulties they would experience compared to L2 speakers and therefore we cannot make any assumptions from our data where we were only able to look at the effect of nativity. As a next step, the current study needs to expand to include participants with hearing loss and elderly participants to investigate how we can make synthetic speech more accessible to them. This will be of particular interest especially to the healthcare arena, where it is vital for robots to communicate intelligibly to their users.

8. Conclusions

Our current findings also support existing robot dialogues such as medicine reminders where medicine names are reported to be hard to comprehend if you hear them by itself without context as shown in (Wolters et al., 2007b; Watson et al., 2013). However the Word Complexity Measure results from this current study suggests that as long as the listeners are familiar with the medicine names and expect to hear them in reminders, the medicine names, being relatively unique phonetically, should be easy to capture.

In this study, we analysed the results from a perception test using semantically unpredictable sentences to examine how different sentence types and word complexities can affect intelligibility in unfamiliar context. While word frequency also affected how well participants could understand the words, phonetic complexity of the word had much greater effects. We also saw nativity as a factor influencing the results, but unfortunately we did not have enough participants to look at age and hearing loss related effects. Overall, we found that in order to make synthetic speech system more intelligible especially for those vulnerable to communication problems such as non-native speakers, dialogue should be designed to have straightforward sentence structures with phonetically unambiguous words.

9. Acknowledgements

We would like to thank our participants who gave us their time to listen to some very strange sentences and to the Centre of Automation and Robotic Engineering and Science, University of Auckland for their support. We would also like to thank our reviewer for their constructive comments.

References

References

Adank, P., Janse, E., 2010. Comprehension of a novel accent by young and older listeners. Psychology and Aging 25 (3), 736–740.

Bartneck, C., Forlizzi, J., 2004. A design-centred framework for social human-robot interaction. RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759), 591–594.
URL http://ieeexplore.ieee.org/document/1374827/

Bates, D., Mächler, M., Bolker, B. M., Walker, S. C., 2015. Fitting Linear Mixed-Effects Models using lme4. Journal of Statistical Software 67 (1), 1–48.

Bauer, L., Warren, P., Bardsley, D., Kennedy, M., Major, G., 2007. New zealand english. Journal of the International Phonetic Association 37 (1), 97–102.

Benoît, C., Grice, M., Hazan, V., 1996. The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. Speech Communication 18 (4), 381–392.
URL http://www.sciencedirect.com/science/article/pii/016763939600026X

Bose, A., Colangelo, A., Buchanan, L., 2011. Effect of phonetic complexity on word reading and repetition in deep dyslexia. Journal of Neurolinguistics 24 (4), 435–444.
URL http://dx.doi.org/10.1016/j.jneuroling.2011.01.004

Brysbaert, M., New, B., 2009. Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior research methods 41 (4), 977–90.
URL http://www.ncbi.nlm.nih.gov/pubmed/19897807

Chen, L.-m., Liu, Y.-h., 2015. The word complexity measure (WCM) in early phonological development : A longitudinal study from birth to three years old. In: Computational Linguistics and Speech Processing. pp. 233–247.

Coalson, G. A., Byrd, C. T., 2016. Phonetic complexity of words immediately following utterance- initial productions in children who stutter. Journal of Fluency Disorder (47), 56–69.

Davies, M., 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. Literary and Linguistic Computing 25 (4), 447–464.

Edlund, J., Gustafson, J., Heldner, M., Hjalmarsson, A., 2008. Towards human-like spoken dialogue systems. Speech Communication 50 (8-9), 630–645.

Ferguson, S. H., Jongman, A., Sereno, J. A., Keum, K. A., 2010. Intelligibility of Foreign-Accented Speech for Older Adults with and without Hearing Loss. Journal of the American Academy of Audiology 21 (3), 153–162.
URL http://openurl.ingenta.com/content/xref?genre=article{&}issn=1050-0545{&}volume=21{&}issue=3{&}spage=153

Fong, T., Thorpe, C., Baur, C., 2001. Collaboration , Dialogue , and Human-Robot Interaction. Robotics (November), 255–266.
URL http://infoscience.epfl.ch/record/29972/files/ISRR01-TF.pdf{%}5Cnhttp://www.springerlink.com/index/45G0U275057L43Q8.pdf

Francis, A. L., Nusbaum, H. C., 1999. Effect of lexical complexity on intelligibility. International Journal of Speech Technology 3 (1), 15–25.

Fraser, M., King, S., 2007. The Blizzard Challenge 2007. SSW.
URL http://festvox.org/blizzard/bc2007/blizzard{_}2007/full{_}papers/blz3{_}001.pdf

Haley, K. L., Jacks, A., 2014. Single-word intelligibility testing in aphasia: Alternate forms reliability, phonetic complexity and word frequency. Aphasiology 28 (February 2015), 320–337.
URL 10.1080/02687038.2013.855702{%}5Cnhttp://offcampus.lib.washington.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true{&}db=a9h{&}AN=93469434{&}site=ehost-live

Hazan, V., Simpson, A., 1998. The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. Speech Communication 24, 211–226.

Hioka, Y., James, J., Watson, C., 2017. A design of comfortable masking sound for real-time informational masking. Internoise 2017, 1283–1291.

Igic, A., Watson, C. I., Stafford, R., Broadbent, E., Jayawardena, C., Macdonald, B., 2010. Perception of synthetic speech with emotion modelling delivered through a robot platform : an initial investigation with older listeners. In: 13th Australasian International conference on Speech Science and Technology. No. December. pp. 189–192.

Jain, S., 2015. Towards the Creation of Customised Synthetic Voices using Hidden Markov Models on a Healthcare Robot. Master thesis, University of Auckland.

Jakielski, K. J., 1998. Motor organization in the acquisition of consonant clusters. Ph.D. thesis, University of Texas at Austin.

King, S., 2014. Measuring a decade of progress in Text-to-Speech. Loquens 1 (1), 2386–2637.
URL http://dx.doi.org/10.3989/loquens.2014.006{%}5Cnhttp:/dx.doi.org/10.3989/loquens.2014.006

King, S., Karaiskos, V., 2016. The Blizzard Challenge 2016. Blizzard Challenge workshop.

Lass, N. J., Dicola, G. A., Beverly, A. S., Barbera, C., Henry, K. G., Badali, M. K., 1979. The effect of phonetic complexity on speaker height and weight identification. Language and Speech1 22 (4), 297–309.

R Core Team, 2015. R: A language and environment for statistical computer. Available at https://www.r-project.org/.
URL https://www.r-project.org/

Schröder, M., Trouvain, J., 2003. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. International Journal of Speech Technology 6 (4), 365–377.
URL http://link.springer.com/article/10.1023/A:1025708916924

Stoel-Gammon, C., 2010. The Word Complexity Measure: description and application to developmental phonology and disorders. Clinical Linguistics & Phonetics 24 (4-5), 271–282.

Watson, C., Liu, W., Macdonald, B., 2013. The Effect of Age and Native Speaker Status on Intelligibility. In: 8th ISCA Speech Synthesis Workshop.

Wolters, M., Campbell, P., DePlacido, C., Liddell, A., Owens, D., 2007a. The Effect of Hearing Loss on the Intelligibility of Synthetic Speech. International Congress of Phonetic Sciences 16 (August), 673–676.

Wolters, M., Campbell, P., Deplacido, C., Liddell, A., Owens, D., Division, A., 2007b. Making Speech Synthesis More Accessible to Older People. In: Sixth ISCA Workshop on Speech Synthesis.