THE UNIVERSITY
of EDINBURGH

THE INTELLIGIBILITY OF SYNTHETIC SPEECH IN NOISE AND REVERBERATION

Karl B Isaac

Thesis submitted for the Degree of Doctor of Philosophy
The University of Edinburgh
2015

*To **all** my children.*
Just because the noise is louder doesn't make it right.

Abstract

Synthetic speech is a valuable means of output, in a range of application contexts, for people with visual, cognitive, or other impairments or for situations were other means are not practicable. Noise and reverberation occur in many of these application contexts and are known to have devastating effects on the intelligibility of natural speech, yet very little was known about the effects on synthetic speech based on unit selection or hidden Markov models.

In this thesis, we put forward an approach for assessing the intelligibility of synthetic and natural speech in noise, reverberation, or a combination of the two. The approach uses an experimental methodology consisting of Amazon Mechanical Turk, Matrix sentences, and noises that approximate the real-world, evaluated with generalized linear mixed models.

The experimental methodologies were assessed against their traditional counterparts and were found to provide a number of additional benefits, whilst maintaining equivalent measures of relative performance. Subsequent experiments were carried out to establish the efficacy of the approach in measuring intelligibility in noise and then reverberation. Finally, the approach was applied to natural speech and the two synthetic speech systems in combinations of noise and reverberation.

We have examine and report on the intelligibility of current synthesis systems in real-life noises and reverberation using techniques that bridge the gap between the audiology and speech synthesis communities and using Amazon Mechanical Turk. In the process, we establish Amazon Mechanical Turk and Matrix sentences as valuable tools in the assessment of synthetic speech intelligibility.

Declaration

I declare that this thesis was composed by me, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Karl B Isaac
March, 2015

Acknowledgements

First and foremost, I should like to thank Steve Renals and Maria Wolters. Not only have they provided the highest level of academic and professional support that could be expected from a PhD supervisor, but also unstinting personal support during some very difficult times.

My time in the Informatics Forum has been made so much easier and enjoyable by the helpfulness and hard work of Avril Heron and her colleagues.

A PhD is never the sole work of one person and many others have contributed in varying degrees, including, but not limited to, the following:

- Members of the HearCom project made available details of their Matrix sentences.
- Wouter A Dreschler kindly donated a CD of the International Collegium of Rehabilitative Audiology (ICRA) noises.
- Vasilis Karaiskos provided access to the data from the *Blizzard Challenge* 2010.
- Heidi Christensen provided the impulse responses, and scripts, from the Computational Hearing in Multisource Environments (CHiME) project.
- The Institute of Communication Systems and Data Processing at Aachen University made available the impulse responses from the Aula Carolina.
- Rob Clark, Simon King, Cassie Mayo, Korin Richmond, Adriana Stan, Cássia Valentini Botinhão, Oliver Watts, and Mirjam Wester provided advice at various times.
- Various authors emailed me copies of papers that were not otherwise available.
- The staff and students of the Informatics Forum, in general, and the Centre for Speech Technology Research (CSTR), in particular, made my time in Edinburgh extremely enjoyable and rewarding.

Last, but not least, this work was made possible by funding from the Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/G060614/1.

Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

GENERAL purpose synthetic speech became generally available as a means of output for computing devices in the early 1980s when the first *DECtalk* unit was released by Digital Equipment Corporation [Friedman, 1983]. Its subsequent deployment in a range of augmentative and alternative communication (AAC) devices and widespread commercial availability meant that it (and the rule-based formant synthesis it was based on) became the focus of numerous academic studies into the intelligibility (and other aspects) of synthetic speech, including in the presence of its most obvious hindrance—noise.

The initial enthusiasm for speech technology waned to the point that Shneiderman and Plaisant [2004, p. 375] described it as, 'the bicycle of user-interface design: It is great fun to use and has an important role, but it can carry only a light load.' However, the increasing eschewal of desktop computers in favour of mobile devices with their small screen and the ubiquity of ear buds mean that two of the arguments against using speech—the superiority of large visual displays and lack of privacy—have fallen away, leading to a resurgence of interest, to the extent that the human-computer interaction (HCI) community is once again hosting workshops dedicated to its use[1].

In the decades since the development of formant synthesis, a number of new systems, using radically different methodologies, have been developed that have improved synthetic speech such that it can now match natural speech for intelligibility in benign acoustic environments (although its naturalness still lags some way behind) [King and Karaiskos, 2009, 2010, 2011].

These newer synthesis systems appear to have received very little attention when

[1] e.g., Designing Speech and Language Interactions (`http://www.cs.toronto.edu/dsli2014/`)

it comes to investigating the effects of noise and reverberation. Indeed, at the inception of the research for this thesis, there appeared only to be one study that had systematically investigated either of the newest text-to-speech (TTS) systems—that is, those based on hidden Markov models (HMMs) or unit selection—at different signal-to-noise ratios (SNRs) [Lancaster et al., 2004] and none that had investigated the effect of reverberation. Yet noise and reverberation remain the Achilles heel of synthetic speech, especially for applications in the home environment, where background noises tend to be many and varied and reverberation ever-present. While van Leeuwen and van Balken [2005] must have tested unit selection systems in their experiment with noise, they do not specify the synthesis methodology used for any of their systems.

If synthetic speech is to achieve its full potential, we need a fuller understanding of the performance of current systems in real world circumstances: the presence of noise and reverberation.

1.1 Goals and contribution

The goal of the research presented in this thesis is the development of a rigorous methodology for the evaluation of speech synthesis and its application to modern speech synthesis systems in a variety of acoustic conditions.

It should be noted that, although in some experiments comparisons with synthetic speech without noise are made, it is not the intention in this work to provide an analysis of the systems in quiet, since this is adequately covered elsewhere (notably in the annual *Blizzard Challenge*, for example [King and Karaiskos, 2012]). Similarly, no analysis or evaluation of naturalness is presented.

As previously mentioned, there is little work currently being carried out that systematically investigates the effects of background noise on modern synthetic speech and none at all on the effects of reverberation. Those studies that do exist [Lancaster et al., 2004; King and Karaiskos, 2010] have tended to rely on evaluations using:

1. undergraduates wearing headphones in a professionally-equipped laboratory
2. semantically unpredictable sentences (SUSS) as stimuli
3. a single, stationary (often unrealistic) noise presented at one fixed SNR
4. stimuli recorded in an anechoic chamber and presented through headphones, that is, conditions that preclude reverberation.

Whilst such methodological decisions may provide experimental rigour, they are not without shortcomings:

1. The concern that results provided by undergraduates using professional equipment in the lab may not be wholly generalizable to the general public in the real world is readily apparent. Less obvious is the possibility that some systems may be better suited, and perform better, outside the lab.

   Whilst not a scientific concern, there is also the practical consideration that engaging participants and funding listening labs are both expensive and time-consuming.

2. Commentators such as Jurafsky and Martin [2008] and Taylor [2009] have raised the concern that tuning TTS systems for SUSS may be counterproductive given that real applications are far more likely to generate sentences that are highly semantically predictable.

   As with using students in the lab, the generalizability of results gained from SUSS, with their inherently higher cognitive load, to older listeners, non-natives, and those with disabilities is questionable. Similarly, SUSS are not used in hearing testing or screening so the potential to correlate synthetic speech intelligibility with hearing (and, therefore, predict it) appears an unnecessarily wasted opportunity.

3. Listeners to synthetic speech probably almost never listen to it in the presence of a single, stationary noise at only one SNR and it has long been known that fluctuating noise is a factor in speech intelligibility [Festen and Plomp, 1990]

4. Likewise, other than when using headphones or ear buds, it is difficult to think of a scenario in which synthetic speech would be heard without reverberation; and it too is known to affect the intelligibility of speech [Whitman, 1915].

This thesis contains the first published evaluation of a crowdsourcing platform—specifically, Amazon Mechanical Turk (AMT)—as an alternative to the laboratory for conducting listening tests. The use of AMT and how it compares with the lab will be of interest to anyone involved in language experiments seeking a cheaper, faster, and more manageable alternative whilst maintaining experimental integrity.

Matrix sentences are proposed as an alternative, and possible solution, to the problems with SUSS. Matrix sentences are meaningful sentences created from a phonemically balanced word set, which was developed for assessing speech intelligibility in noise as part of an audiological test battery in the HearCom project [Wagener, 2009]. The analysis of Matrix sentences will provide a bridge to audiological research not provided by SUSS that may allow audiologists to predict a person's ability to understand various types of synthetic speech given the results of an audiological test.

Realistic, non-stationary background noises that have been carefully studied and validated are added to natural and synthetic speech before being presented at a number of levels likely to be representative of those occurring 'in the wild' and to have a significant effect on intelligibility.

The effects of reverberation are investigated for the first time as are the combined effects of noise and reverberation. The research into the effects of noise and/or reverberation will pinpoint problematic SNRs and levels of reverberation for the two currently prevalent TTS systems, the HMM-based Speech Synthesis System (HTS) and *Festival* unit-selection (F-USEL), thus allowing for their eventual improvement and that of their successors.

1.2 Thesis outline

The remainder of this thesis is structured as follows.

Chapter 2: Evaluating the intelligibility of synthetic speech provides a comprehensive literature review, which introduces the main concepts that form the basis of the rest of the thesis. Background information is provided on: how speech is synthesized, focusing on the different methods that have been put forward; work in the field of human speech research, particularly tests of intelligibility in quiet and in noise and the types and levels of noise used in testing; and a review of the intelligibility of modern synthetic speech in noise.

Chapter 3: Experimental methodology sets out the recent developments in other fields that were identified that could be used for the assessment of synthetic speech and an evaluation of their efficacy compared with existing standard techniques. These newer techniques include AMT, Matrix sentences, and an ecologically-valid background noise.

Chapter 4: Synthetic speech in noise includes an overview of the effect of noise on human speech. It identifies the effect that various noises have on natural speech and two synthetic versions, taking into account other factors, such as the age and hearing ability of the listener. The chapter is underpinned by, and makes use of, the evaluation of the techniques proposed in Chapter 3.

Chapter 5: Synthetic speech in reverberation reviews the causes of reverberation and its effects on natural and synthetic speech, in much the same way as noise was investigated in Chapter 4, once again calling on the work presented in Chapter 3.

Chapter 6: Synthetic speech in noise and reverberation. Since the effect of noise and reverberation is greater than the individual components [Nabelek and Mason, 1981; Payton et al., 1994] it is important to study their combined effect on synthetic speech. Chapter 6 brings together the work of the preceding chapters and shows how

intelligibility is affected by this combination.

Chapter 7: Conclusion and future work summarizes the thesis and evaluates its contribution to the wider speech synthesis community. Longer term goals, perhaps for post-doctoral work, are set out for future work that could not be completed during the PhD time frame and for ideas that arose from the work carried out.

CHAPTER 2

EVALUATING THE INTELLIGIBILITY OF SYNTHETIC SPEECH

SPEECH output can be provided in one of two ways: recording a human voice and storing the samples as digitized sound waves; or using a computer to generate digitized sound waves based on an algorithm, model, or a set of rules. The former is known as digitized (or canned) speech and the latter as synthetic speech. Canned speech has the advantage that it can sound more natural, but is generally limited to a few stock words or phrases. Synthetic speech, on the other hand, is generally able to synthesize any valid utterance in the target language [Taylor, 2009]. The preference for most applications would be synthetic speech so that names of people, drugs, and other out-of-vocabulary words can be used and the message can be personalized for individual users.

2.1 Synthetic speech

The most common form of synthetic speech production involves the transformation of strings of text to audible acoustic waveforms and, consequently, is often known as text-to-speech (TTS). Other production methods can include those that use concepts or phonemes as input. Whilst the transformation of text to speech might sound straightforward, in practice it is not and involves: determining *what* needs to be said; determining *how* it needs to be said; making an internal representation of the utterance; and generating a waveform from the internal representation.

The first step is sentence tokenization, that is, establishing where each sentence

begins and ends, whilst accounting for end-of-sentence markers, such as, exclamation marks, full stops, and so on being used *within* a sentence, for example in abbreviations.

The next step is to normalize non-standard text, in other words, convert various shorthands, such as numeric dates, pound signs, and abbreviations to their expanded version as they would actually be spoken by a human.

Once the sentences have been determined and the words that need to be spoken identified, work can begin on pronunciation. Firstly, words that have multiple pronunciations must be identified, such as the present and past verb forms of the word 'read'. For the English language, the actual pronunciation of the words can be provided by referring to a dictionary of pronunciations. Of course, this will only work when the word to be pronounced appears in the dictionary. The pronunciation of out-of-dictionary words has to be established using a letter-to-sound system based on statistical likelihood or neural networks.

After the pronunciation of the individual words has been calculated, algorithms to ascertain the intonational and rhythmic properties (or *prosody*) of the utterance are employed. The prosody of a spoken utterance can provide information about meaning (for example, whether the utterance is a statement or question), emotion, social context, and other paralinguistic features.

For most synthesis methods, other than unit selection, which is described below, the fundamental frequency (F0) and duration have to be calculated for each segment of speech. At this point, the system now has an internal representation of all the information it needs to generate the waveform and it is at this stage that the various synthesis methods tend to diverge. Today's most commonly encountered systems—articulatory, formant, diphone, unit selection and hidden Markov model (HMM)—are described below together with a discussion of their relative qualities.

2.1.1 Articulatory synthesis

The very earliest attempts at synthesizing speech were based on modelling the human articulatory system, the first published account of which is that of von Kempelen [1791]. In the past, the technique suffered from the difficulty of discovering exactly how the human body produces speech and, therefore, what parameters to pass to any model of it. The technique has not played a large part in mainstream speech synthesis and is not, therefore, covered further here. However, in recent years the advent of techniques such as magnetic resonance imaging and electromagnetic articulography has enabled the use of articulatory features—often in specialized contexts—in the control of some of the systems described below.

2.1.2 Formant synthesis

The name 'formant synthesis' derives from the fact that the method uses formant information to generate waveforms for voiced speech. Because it does this by following a set of rules, it is also known as *synthesis by rule*, although all synthesis techniques employ rules of one kind or another.

The technique was used by Holmes et al. [1964] to copy synthesize sentences that were, eventually, virtually indistinguishable from the originals. It is the technology used in the very popular *DECtalk* system mentioned above, which was based on *MITalk* [Allen et al., 1987].

Formants are generated by a model that is a (very) rough approximation of the vocal tract, usually comprising components that model the nasal and oral cavities. A periodic signal (for voiced sounds) and/or white noise (for obstruents) can be passed through the oral-cavity components and/or the nasal-cavity components (for nasalized sounds) and then through a 'radiation' component that emulates the effect of the lips and nose.

One of the advantages of formant synthesis is that it provides almost complete control over the input parameters, such as formant values, duration, and F0. Having said that, the basic nature of the model means that—on its own—the naturalness of its output rarely approaches that of human speech. Perhaps because of this, though, and the fact that the targets are clear and distinct from one another, even if the transitions between them are unnatural, formant synthesis has proved to be very intelligible [Taylor, 2009]. Mirenda and Beukelman [1987] even found *DECtalk* as intelligible as human speech. In fact, partly because it suffers from fewer signal processing artefacts than other systems, it remains intelligible even at high speeds, making it invaluable in applications—such as screen readers—for people with visual impairments. Its relatively low requirements for memory and storage mean that it can be deployed on devices with limited capabilities and is, therefore, particularly useful for augmentative and alternative communication (AAC) devices.

2.1.3 Diphone synthesis

Diphone synthesis is one of a family of concatenative techniques that takes parts of pre-recorded speech and joins the parts together to produce the desired output. A diphone is created by cutting a piece of speech extending from the middle of one phone to the middle of the following phone. The diphones are then joined together to make up the waveform described by the internal representation. Diphones are used in preference to phones to avoid discontinuities that may occur because of coarticulation. Coarticulation is the assimilation of moving parts of the articulatory system (tongue,

lips, etc.) either in preparation for saying the next phone or as a result of saying the previous phone. Since there is, generally, more movement of the articulators (and, hence, more variation) *between* phones, better joins can be achieved by joining at the *middles* of phones. However, just using diphones does not exclude all discontinuities in the joins and some signal processing is required to smooth the transitions between them. Moreover, the pitch, energy, and duration of the concatenated diphones is unlikely to match that required and will need adjusting. Whilst the rudiments of a diphone synthesis system had been demonstrated by Olive [1977], its success hinged upon finding techniques to minimize distortions in the waveform whilst generating the correct prosody and maintaining naturalness. The Pitch-Synchronous and OverLap-Add (PSOLA) techniques of Moulines and Charpentier [1990] were instrumental in this regard. Further improvements were made with the development of the Multiband Resynthesis OverLap-Add (MBROLA) techniques of Dutoit and Leich [1993].

Even with the additional signal processing, the final result will contain some artefacts caused either by the signal processing itself, or the fact that diphones are created only with consideration of their nearest neighbours, ignoring the fact that other phones further away may be exerting an influence.

Although, relative to model- or rule-based systems, diphone synthesis may be seen as a data-driven approach to the problem of speech synthesis, its data storage requirements are relatively small, since only one example of each diphone type needs to be stored. However, this does mean that the quality of the synthesis is largely dependent on the quality of the stored diphones, which, in turn, means that great care should be taken in obtaining the diphones from speech.

2.1.4 Unit selection synthesis

Unit selection synthesis is similar to diphone synthesis in that it is a concatenative system. However, instead of only concatenating diphones, it is capable of concatenating any arbitrary length *unit*. In order to maximize the benefits of unit selection, as large a database as possible of speech units is created, each with an extensive set of linguistic features that differentiate it from similar units and can be used as a basis for selection. In this way, it is possible that whole phrases or sentences of the desired utterance will exist in the database and can be selected and used with very little further processing. When nothing of the desired utterance exists in the database, diphone synthesis can be used as a fall-back position. The first unit selection system was written for the Japanese language [Sagisaka et al., 1992] before being used for English in the *CHATR* speech synthesis system [Black and Taylor, 1994] that was specifically designed to

facilitate speech synthesis research.

Using very large databases of speech units raises the problem of deciding which unit will provide the best output. A solution was proposed by Hunt and Black [1996] that eventually lead to unit selection becoming the pre-eminent synthesis method. Their solution is based on assigning a target cost and a join cost to each possible unit and selecting the one with the minimum costs. The target cost is a measure of how different the actual linguistic features of the unit are from those required and the join cost is a measure of how far the join between one unit and the next is from being ideal. (Bear in mind that the units in the database are as diverse as possible and are not selected for their ability to join with their neighbours, as is the case with diphone synthesis.) The best sequence of unit $\hat{U}$ is then given by:

$$\hat{U} = \arg\min_{u} \left\{ \sum_{t=1}^{N} T(u_t, s_t) + \sum_{t=1}^{N-1} J(u_t, u_{t+1}) \right\} \tag{2.1}$$

where u_t is a unit, s_t is a set of desired features, T is the target cost, and J the join.

Unfortunately, if a suitable unit is not available, the system fails to produce acceptable output in a rather obvious fashion. Other disadvantages of unit selection synthesis tend to come from its need for very large databases, the lack of control, and lack of expressiveness. The most obvious, the physical storage required, may not be a problem on modern desktop computer systems, but still poses problems for mobile and more limited devices. Another is the need to record all the features of each unit, which, as we have seen, includes target and two sets of join features (one for left joins and one for right). Recording the features is either a laborious hand-labelling task or a machine-learning problem, neither of which is easy or fast. Fortunately, this only has to be done once. The advantage of using a large database—the ability to produce near-perfect output when the desired units are in the database—can also be a double-edged sword in that the output is perceived as inconsistent when the desired units are not found.

2.1.5 Hidden Markov model-based synthesis

Hidden Markov models were first used in automatic speech recognition (ASR) rather than speech synthesis [Baker, 1975]. In ASR, speech is captured in a series of frames, each with a set of characteristics of the acoustic information it holds and models of phones consisting of multiple states are built. From a set of training data, it is possible to build up a model comprising a sequence of states where each state models a sequence of frames. When trying to recognize an utterance, the most likely state sequence is found that would generate the speech in question. The process

can be further refined and enhanced by storing the rate of change of the acoustic characteristics between frames (the velocity or delta coefficients) and the rate of change of those changes (the acceleration or delta-delta coefficients). Each phone model is further divided into a number of states, representing different parts of the phone. The probability of moving between each state, and back to itself, is modelled so that it is possible to discover which sequence of states is most likely to have given rise to the series of frames making up the phone and, hence, the series of phones making up the utterance.

Synthesizing speech using HMMs would seem to be, essentially, the reverse of recognition and early work showed this to be the case [Ljolje and Fallside, 1986; Falaschi et al., 1989]; however, it was the seminal work of Tokuda et al. [1995] that lead the way for HMM-based synthesis. Speech synthesis differs from recognition in that it uses the HMMs to actually generate parameters and has four fundamental requirements. Firstly, the most likely sequence of states and frames is generated from the phone models usually taking into account the velocity and acceleration coefficients to ensure a smoother waveform is created. Secondly, there are multiple streams of parameters for F0 and prosodic features that are important for natural sounding speech. Typically, there will be more than 50 parameters describing the periodic and aperiodic spectral envelopes and the value of F0. Thirdly, duration modelling has to be carried out separately, meaning that most HMM-based synthesis actually uses a hidden semi-Markov model (HSMM). Finally, a model of the vocal source is required to provide the actual output, which needs to be able to accept the parameters generated by the HMM process. A full explanation of the techniques now used in HMM-based synthesis can be found in the review by Zen et al. [2009].

Although HMM synthesis, like unit selection, can be described as a data-driven approach, unlike unit selection, it does not generate speech directly from the data, but from a statistical model of the characteristics of speech previously learnt from the data. Thus, HMM-based synthesis systems require far less space for data storage than unit selection and so can be deployed on devices with very limited capabilities. Because building speech models is a statistical process that identifies the parameters that are used to generate the speech, it does not require a large speech corpus, although, generally speaking, performance is improved if one is used. When a large set of data is available, an average voice model can be built, which can then be adapted from a smaller set of data to change the characteristics of the output, so that it can be made to sound like a particular person [Creer et al., 2009], or emulate emotional speech or various speaking styles [Tachibana et al., 2005].

One of the disadvantages of HMM-based synthesis is that, because it does not use recordings of real speech directly, the quality of the output is dependent on the quality

of the source model, which can lead to its sounding buzzy, although using mixed excitation can alleviate this.

Another disadvantage is that the speech can sound muffled because of the use of statistical averaging in the parameter generation of the trajectory model. Again, this can be alleviated using a technique called *Global Variance* [Toda and Tokuda, 2005], although this can generate its own problems, particularly artefacts in short utterances.

2.1.6 The quality of synthetic speech

What constitutes 'good' synthetic speech is not clearly defined in the literature, although most commentators will separate out 'quality' and 'intelligibility', both of which are usually assessed by human listeners [Loizou, 2007; Jurafsky and Martin, 2008; Taylor, 2009].

Synthetic speech quality is difficult to measure as it is a subjective measure and, although it may seem clearly defined as, say, 'naturalness', it is often not made clear what constitutes naturalness. The focus of the work presented here is on evaluating the intelligibility of synthetic speech (particularly in noise) rather than its quality. Fortunately, intelligibility is a more objective measure, as participants report what they actually hear, and is relatively straightforward to measure having much in common with the measurement of the intelligibility of natural speech.

2.2 Evaluation of speech intelligibility

Intelligibility may be defined as 'the proportion of speech items (e.g., syllables, words, or sentences) correctly repeated by (a) listener(s) for a given speech intelligibility test' [Brand, 2009, p. 197]. However, the differing needs of researchers with an interest in measuring intelligibility, such as audiologists, speech scientists, and engineers, have, over the years, led to a wide range of intelligibility tests being proposed at the syllable, word, and sentence levels; the more common of which are presented below. Of these measures, many have been designed to test the intelligibility of speech in quiet and, whilst some may also be used with noise—either specifically to test intelligibility in noise, or to make the test more discriminative—only a few have been *especially* designed to test the intelligibility of speech in noise.

It should be noted here that some sources (for example, Allen [2005]) differentiate between intelligibility at the segmental level, where there is no meaning to be understood, and the intelligibility of meaningful sentences, where 'intelligibility' includes comprehension of the meaning; however, the term is widely used in the literature

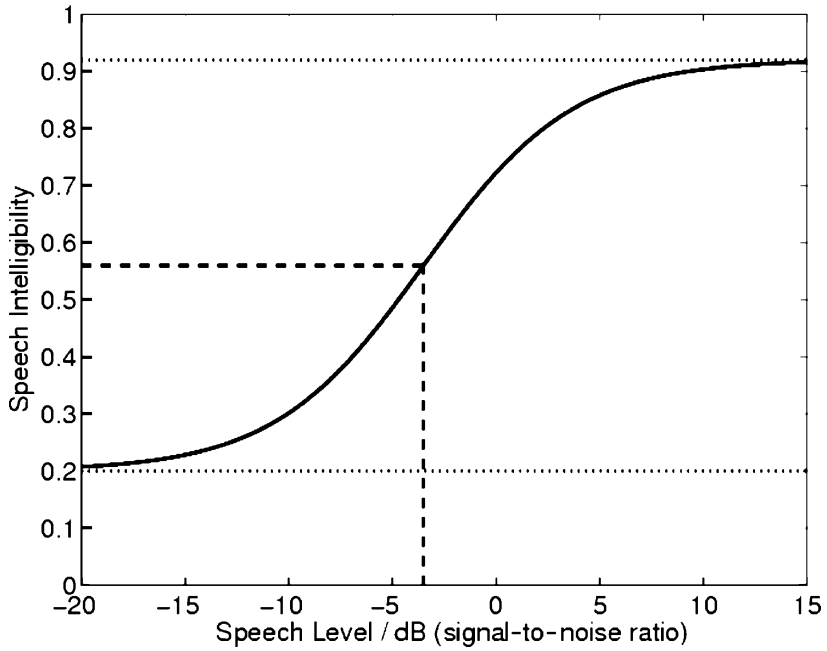

Figure 2.1: *Typical example of speech intelligibility function curve (solid line). Reproduced, with permission, from Brand [2009]. In this particular example, the dotted lines denote the upper and lower bounds of intelligibility in practice and, hence, the SRT occurs nearer 60 % of the theoretical maximum intelligibility*

to cover both meanings. In most cases, where the materials are presented at a fixed speech level or signal-to-noise ratio (SNR), the proportion of correctly identified speech items (converted to a percentage) is used as the measure of intelligibility. In circumstances where it is necessary to differentiate between two treatments, though, this may result in ceiling effects that make it difficult to differentiate between them.

An alternative approach is to vary the speech level, or SNR, systematically, from a point where all the speech can be understood to where none of it can (or vice versa). The resulting data points represent an empirical manifestation of the intelligibility function—that is, the listener's speech intelligibility, as a function of the speech level or SNR in decibels (dB) [Brand, 2009].

In practice, the range of intelligibility may be limited, for example, by distortions in the speech stimuli, such that the minimum or maximum theoretical intelligibility can never be achieved. The dB level at which 50 % intelligibility occurs is known as the speech reception threshold (SRT). As psychoacoustic research has shown, illustrated in Figure 2.1, this point is the steepest part of the curve and, therefore, yields the highest sensitivity.

Instead of taking measurements along the whole of the speech intelligibility curve, which is laborious and tedious for both tester and listener, finding the SRT may be simplified by using one of the *up-down* procedures described by Levitt and Rabiner

[1967] and Levitt [1971], or similar variants. These procedures are an efficient and practical way of finding the 50 % level in far fewer measurements. The basic premise is that the SRT can be found by adaptively increasing or decreasing (using a given step size) the SNR at which stimuli are presented in response to incorrect and correct answers, respectively. The stimuli usually consist of a fixed set of words or sentences of equal difficulty. The SRT is then estimated by taking an average of a fixed number of the last stimuli presented. Alternatively, a sequence of changes of SNR in the same direction is identified as a run the midpoint of which is used as the estimate. The size of the increase or decrease in SNR (the step size) and whether or not it varies during the procedure is one source of a variety of enhancements that have been proposed.

2.2.1 Syllable-level tests

Syllable-level tests consist of nonsense syllables formed from a particular combination of consonants and vowels. Early tests consisted of more or less random arrangements of consonant-vowel, vowel-consonant, or consonant-vowel-consonant (CVC). Standardization and refinement by Fletcher and Steinberg [1930] led to the common use of CVC nonsense syllables in intelligibility testing. They also introduced the concept of introductory sentences to be said before the syllable in order to bring its articulation closer to that of everyday speech [Fletcher and Steinberg, 1930].

Miller and Nicely [1955] used a modified syllable test to establish which—and to what extent—English consonants were confused with one another and what features of these consonants were associated with the confusion.

Some criticisms of syllable-level tests are that: it can be difficult to construct lists of equal difficulty; some training of listeners may be required [Loizou, 2007]; and they do not truly reflect the intelligibility that the speech would have in an everyday sentence.

In turn, one of the advantages of nonsense syllables is that there is no context to confound the intelligibility score [Fletcher and Steinberg, 1930]. They may also be particularly useful when investigating the intelligibility of some specific aspect of language in isolation, such as, consonants.

2.2.2 Word-level tests

Word-level tests were devised in order to overcome some of the perceived problems of nonsense syllables.

Egan [1948] developed 20 sets of 50 common English words (often known as the Harvard test). The word lists were phonemically balanced so that they represented the distribution of sounds in the English language and may also be known as the

phonetically-balanced (PB) test for this reason. (Exact correspondence to the language was not possible because of the need to meet other requirements for the lists, such as, that the lists be of equal difficulty and that the words themselves should be of similar difficulty within each list, to maximize discriminability.) Egan states that his word lists do not require extensive training of listeners, unlike when using nonsense syllables, although he does recommend training listeners until no further improvement results, especially under difficult listening conditions, to obviate learning effects [Egan, 1948].

Other word-level tests are based on rhyming words that differ only in the initial or final consonant. Fairbanks [1958] developed 50 sets of five monosyllable, CVC, rhyming words, in which the words were also spelt the same, apart from the initial consonant. Considerable effort was employed in making sure that the words were well known, reflected the language, and were not biased against by taboo, spelling, or variations in pronunciation. The test was known as the Rhyme Test and has formed the basis of a number of variations. The Modified Rhyme Test (MRT) [House et al., 1965] is a variation on the Rhyme Test using (originally) closed sets of six words that differ in their final consonants.

Another variation, the Diagnostic Rhyme Test (DRT) [Voiers, 1983] is still widely used today. It is similar to the MRT, but uses sets of two words and the concept of *phonemic features* proposed by Miller and Nicely [1955]. As its name suggests, in addition to providing a score of intelligibility, it gives a 'diagnosis' of which features of the language are causing confusion. The DRT was shown to be reliable and to be useful in speech-in-noise conditions [Voiers, 1983], although Smith [1979] found that using different talkers produced highly significant differences in DRT scores.

Word-level tests are presented either as individual words or as part of a neutral carrier sentence and, therefore, provide no context to the listener, apart from the word itself. However, it could be argued that in the real world, little, if any, communication is conducted without some form of context and a number of sentence-level tests have been created to accommodate this fact.

2.2.3 Sentence-level tests

One of the first sentence-level tests consisted of 100 phonetically-balanced sentences and became known as the Harvard Psychoacoustic Sentence Test [Egan, 1948]. The sentences are syntactically and semantically normal, which means they can suffer from learning and ceiling effects, but are easy to administer and contain normal prosody.

Whilst sentence-level tests may be seen as being more representative of real-world speech, the longer the utterance used, the more difficult it is to ensure equality within

and between tests. This fact was demonstrated by the development of the Speech Perception In Noise (SPIN) test by Kalikow et al. [1977]. The authors went to great pains to develop a set of sentence-level test materials that were representative of the language in terms of word frequency and phonemic balance and that were balanced for length, intelligibility, and key-word predictability. Despite this, however, later work by Morgan et al. [1981] cast doubt on the equivalence of the forms of the test and criticized the fact that they had only been tested with people with normal hearing.

Although the SPIN test is a sentence-level test, it actually relies on the respondent identifying the last word of the sentences. The sentences themselves fall into two classes: high predictability or low predictability. High-predictability sentences are designed to make the final (test) word more predictable by leading the listener to it. For example, the sentence, 'The boat sailed across the bay', makes the word 'bay' very predictable; whereas, it is not predictable from the sentence, 'John was talking about the bay.' Scoring of the test uses the high and low predictability sentences to control for the level of semantic information.

An alternative to the SPIN test is the Hearing In Noise Test (HINT) developed by Nilsson et al. [1994]. This test consists of 25 lists of 10 (American) English sentences and is scored on the percentage of correct words from the whole sentence. It was designed to be presented using an adaptive SRT method, that is, the presentation level is increased for incorrect responses and decreased otherwise, as described above. The mean of the fifth and subsequent presentation levels is used as the SRT. This is in contrast to the SPIN test, which is presented at a fixed SNR, typically of 8 dB.

Killion et al. [2004] identified some potential disadvantages of the HINT test, namely that it is based on whole-sentence scoring, so more sentences are required to achieve statistical reliability than when using key-word scoring, and that it uses a stationary noise, which is not as realistic as babble noise. However, they did acknowledge that these aspects could be seen as advantages, depending on the requirements of the test.

In order to overcome these perceived disadvantages and to formulate a speech-in-noise test that could be used as part of a battery of audiological tests, they developed the Quick Speech In Noise (QUICKSIN) test [Killion et al., 2004]. To facilitate its widespread use in audiology, they recognized that it would need to be quick, easy to administer, simple to score, be seen to have validity, and be useful for those with or without hearing impairment.

QUICKSIN is based on the earlier speech in noise (SIN) test [Killion and Villchur, 1993], which experienced a number of problems, not least of which was the fact that audiologists reported its taking too long to administer [Killion et al., 2004]. The test

consists of 12 equivalent lists of six Institute of Electrical and Electronics Engineers (IEEE) sentences [IEEE, 1969] that can be used with normal and hearing-impaired listeners, presented in a four-talker babble noise at SNRs from 25 dB to 0 dB in 5 dB steps. (A further six lists are also available.) It is scored on the number of target words (five per sentence) correctly identified. The QuickSIN is unusual in that it uses a female speaker for its stimuli.

The use of a female talker apparently made the QuickSIN test too difficult for some cochlear implant patients and a revised test, the BKB-SIN test, was devised as a result [Niquette et al., 2003]. It is called the BKB-SIN test because it uses the Bamford-Kowal-Bench sentences originally developed for use with partially-hearing children [Bench et al., 1979]. The test consists of 18 list pairs, the first eight of which consist of ten such sentences recorded at SNRs from 21 dB to −6 dB in 3 dB steps by a male speaker and presented in the same four-talker babble as the QuickSIN (the remaining ten list pairs are for use with cochlear implant patients). The Bamford-Kowal-Bench (BKB) sentences used in the BKB-SIN test have more context than the IEEE ones used for the QuickSIN. The BKB-SIN test is scored on the number of keywords (four for the first sentence; three for subsequent sentences) identified correctly.

Both the QuickSIN and the BKB-SIN tests are presented using a descending level paradigm, that is, the sentences are presented at an increasingly unfavourable SNR at steps determined by the test protocol. In a descending-level method the 50 % point can be calculated using the Spearman–Kärber formula [Finney, 1952], the general form of which is given by:

$$50\,\% = i + \frac{1}{2}(d) - (d)(n)/(w) \tag{2.2}$$

where i is the initial presentation level; d is the decrement size in dB; n is the number of words correct; and w is the number of words per decrement. Although, in practice, this can be simplified to Equation (2.3) for the QuickSIN test [Killion et al., 2004] and Equation (2.4) for the BKB-SIN [Niquette et al., 2003], which has four keywords in the first sentence.

$$50\,\% = 27.5 - (n) \tag{2.3}$$

$$50\,\% = 23.5 - (n) \tag{2.4}$$

An evaluation of the BKB-SIN, HINT, and QuickSIN tests, along with Wilson's own Words In Noise (WIN) test [Wilson, 2003] can be found in Wilson et al. [2007].

Generally speaking, one would expect sentence-level tests to take longer to administer than word- or phone-level tests, but using descending-level or adaptive presentation methods can result in a test that can be administered very quickly. With

the (adaptive) HINT test, 'A threshold measurement with a single list usually takes less than two minutes' [Nilsson et al., 1994, p. 1095] and with the (descending) BKB-SIN test, 'Each list of 10 sentences required 95 s to administer.' [Wilson et al., 2007, p. 847]. A further advantage of SRT-based tests is that they avoid ceiling and floor effects. Moreover, when assessing the speech perception of listeners with hearing aids, using stimuli less than sentence length may not give the hearing aid sufficient time to carry out the necessary processing [Nilsson et al., 1994].

A potential disadvantage of sentence-level tests is that they may suffer from memory effects, since the respondent is expected to remember the whole sentence—rather than a single word—before responding. This is ameliorated to some extent in the HINT test by scoring slight variations as correct, for example substituting 'a' for 'the' or a past-tense verb for the present tense and in the QUICKSIN and BKB-SIN tests by focusing on key-words. Potentially, it is more of a problem in tests based on the semantically unpredictable sentence (SUS).

Although Miller and Isard [1963] first used *semantically anomalous* sentences, as they called them, to investigate the role of syntactic and semantic rules in the perception of sentences in natural speech, almost without exception [van Wijngaarden et al., 2002], SUSs have become a test of *synthetic* speech and will be discussed in detail in Section 2.4.

A potential alternative to SUSs are Matrix sentences, which were developed for assessing speech intelligibility in noise as part of an audiological test battery for the HearCom project[1] [Wagener, 2009]. Matrix sentences are comparable to Oldenburg sentences [Kollmeier and Wesselkamp, 1997] as both are derived from the sentences created by Hagerman [1982]. Each sentence consists of exactly five words in the form 'Name verb quantity adjective noun', such as 'Barry likes nine small spoons'. Sentences are created by randomly choosing from a total of 50 words, ten from each of the five word categories. The fifty words were chosen to be balanced and to cover the phoneme set of British English. The resultant sentences have meaning (and are, therefore, more memorable), but the individual words are relatively unpredictable, thus reducing the potential advantage of contextual information.

2.2.4 The use of noise in intelligibility testing

Although intelligibility tests may be presented to listeners in quiet at a standard loudness or sound pressure level (SPL), they are often presented at various SPLs or SNRs and in one or more background noises. The variation of the SPL or SNR is usually carried out to establish the SRT. The use of a background noise might be to facilitate

[1] http://hearcom.eu

variation of the SNR or more specifically; to mask part, or all, of the speech sound wave; or, more pragmatically, to test the impact of a specific noise on intelligibility.

Varying the SPL or SNR systematically from high to low (or vice versa) has the effect of sensitizing the test, that is, making it more sensitive to small variations in intelligibility and is often done as part of a SRT method of testing. By reducing the SPL or SNR the speech becomes more difficult to hear and deficiencies in the speaker's intelligibility or the listener's hearing become apparent.

2.2.5 Types of noise used in intelligibility testing

One of the earliest uses of noise in testing was as a mask in audiometry. Here, the intention is to restrict the non-test ear's ability to contribute to the measurement for the test ear. Audiologists recognize three types of noise: broadband, speech spectrum, and narrowband; or white noise, pink noise, and narrowband masking noise respectively.

White noise is, perhaps, the most obvious choice of masking noise since it contains a range of frequencies at a constant SPL at each frequency and would be expected to mask any and all test frequencies. It has been used in intelligibility testing since the 1940s [Licklider and Miller, 1948]. However, it was discovered by Fletcher [1940] that once a critical bandwidth around the frequency to be masked had been applied, applying frequencies outside the critical bandwidth only added to the overall level of noise without additional masking of the target frequency. White noise is also less intense at lower frequencies.

Pink noise provides more efficient masking of speech since it has equal energy in each critical band for normal-hearing listeners. Having an energy content that decreases with frequency at -3 dB per octave, pink noise acts as a white noise filtered to resemble the speech spectrum. It has more energy than white noise in the low frequencies.

Narrowband noises may be created with the desired critical bandwidth by passing a broadband noise through narrow band filters. In audiology, these are used to mask the pure tones used in hearing tests.

In addition to white, pink, and narrowband noises, intelligibility tests may use a range of other noises, many of which are speech-related. The most common is the multi-talker babble, that is, between one and one hundred people talking at the same time. In this case an actual recording may be made of several people talking at once or a simulation made electronically from one or more people talking.

Speech-related noises may be presented as modulated or unmodulated, that is, with or without the sorts of pauses and variations that occur in real speech.

In specific circumstances, noises from a particular domain, such as cockpit, machine gun, factory, or office noises may be used.

2.2.6 Levels of noise used in intelligibility testing

Sound is carried through the air as a wave, which may be expressed in terms of pressure in Pa. The human ear is able to hear in the range of 20 μPa (just audible to a young, healthy ear) to 20 Pa (the point at which pain occurs). Since this results in an impractically large scale, a logarithmic scale of dB SPL is used instead, defined as:

$$\text{dB SPL} = 10 \log_{10} \left(\frac{p}{p_0} \right) \tag{2.5}$$

where p is the sound measured and p_0 is the reference value, 20 μPa.

Noise levels are usually expressed as the SNR in decibels. Since the decibel scale is logarithmic, an SNR of 0 dB occurs when the SPL of the signal and the noise are equal and the SNR rises by about 6 dB for every doubling of the signal's SPL and decreases by 6 dB for every halving of the signal's SPL relative to the noise.

Most human speech takes place at a level of 40 dB to 60 dB SPL up to a maximum of approximately 65 dB SPL measured at one metre from the speaker [Fant, 2005], so it is common practice to normalize all speech samples to a level within this range and to apply noise relative to that level. Deliberately applying a noise to a speech signal for testing purposes may be achieved by one of two (or both) methods: additive and multiplicative [Rothauser et al., 1968].

The additive method simply adds the noise signal to the speech signal in much the same way as a background noise in a room gets added to the speech of a talker in that room. Periods of speech silence will become periods of noise (assuming the noise signal does not happen to be silent at that period).

The multiplicative method applies noise by convolving the speech signal with the noise signal, so that any periods of speech silence will remain silent, and reflects the effect of channel noise. Which method is chosen depends on the purpose of the test and may have implications for the accuracy of the SNR calculation.

Speech-in-noise stimuli, theoretically, can be presented at any SNR or SPL level; however, exposure to noise can cause damage to the auditory system, with the amount of damage caused being, primarily, a function of the intensity of the noise and its duration [Katz et al., 2009]. Most guidelines on hearing protection use a 3 dB exchange rate rule, that is, halving the length of allowed exposure for every 3 dBA rise in noise above 90 dBA over the working day (normally defined as eight hours).

The SNR chosen for a particular test will depend on the requirements for that

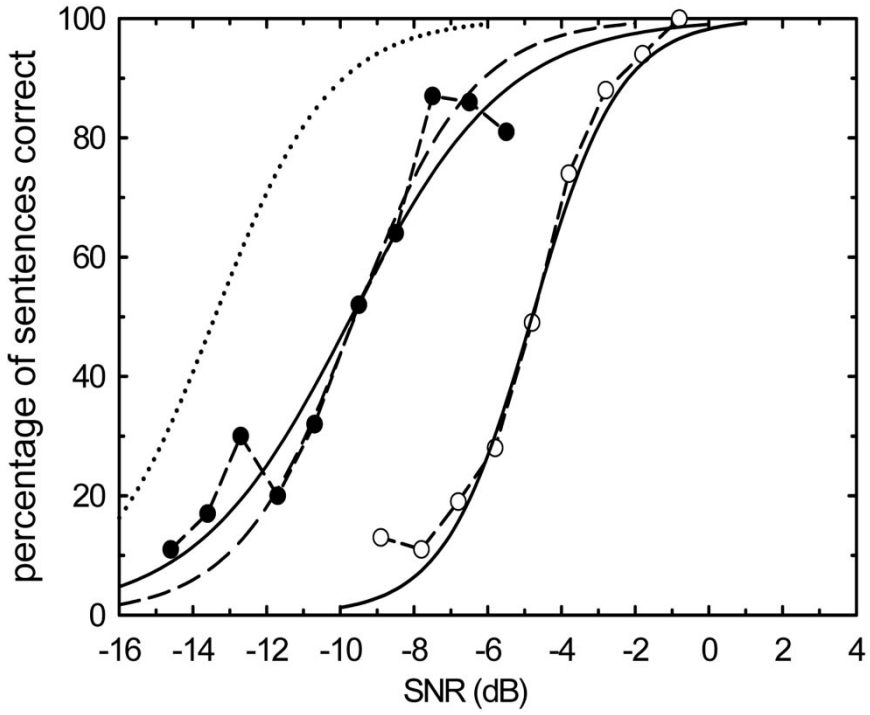

Figure 2.2: *Percentage of sentences correct as a function of* SNR *(dB), for a stationary noise masker (open symbols) and fluctuating noise masker (filled symbols) (replotted from Festen and Plomp [1990]). The two solid curves represent Festen and Plomp's fit to the data. The dotted curve is predicted by the extended* SII *model, based on the curve given by Festen and Plomp [1990] for stationary noise. The dashed curve (without symbols) is identical to the dotted curve, except for a 3.8 dB shift to the right. Reproduced with permission from Rhebergen, K. S. and Versfeld, N. J. (2005). A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners.* The Journal of the Acoustical Society of America*, 117:2181. Copyright 2005, Acoustical Society of America.*

test. If the aim is to assess the effect of a particular level of a background noise, then the stimuli may be presented at an appropriate, fixed SNR. When establishing the SRT, the stimuli will be presented at systematically varied SNRs so that the level at which 50 % of speech is intelligible can be ascertained. In the highly standardized, generic, stationary noises such as white noise, pink noise, and multi-talker babble that speech intelligibility is typically assessed, the SRT occurs at about −4.5 dB, whereas in fluctuating noise it occurs at about −12 dB as demonstrated in Figure 2.2 [Rhebergen and Versfeld, 2005].

2.2.7 Summary of intelligibility tests

Table 2.1 provides a summary of some of the intelligibility tests available, from which it can be seen that measuring the intelligibility of speech in noise is a complex process of choosing a test, scoring system, analysis system, type and level of background noise even without the added complication of having to use human listeners to

carry out the assessment. It would be much simpler if it were possible to *predict* the intelligibility from the speech and noise directly.

2.3 Predicting speech intelligibility in noise

Predicting speech intelligibility in noise is dependent on being able to create a model of the speech, the noise, and the hearing ability of the listener and having a transfer function that can convert this to intelligibility. The factors that govern intelligibility and constitute such a model were first reported by French and Steinberg [1947]. Since that time, our understanding has developed and been refined to the point where a Speech Intelligibility Index (SII)—formerly the Articulation Index (AI)—can be calculated that specifies how much of the audible speech is available at the listener's ear. The method for calculating the SII is fully specified by ANSI [2007]. Essentially, it involves calculating the average speech spectrum, the average noise spectrum, and the listener's hearing threshold and using them to produce an index between zero and unity. This is achieved by: *a*) filtering the speech and noise spectra into bands; *b*) making adjustments for the upward spread of masking, inaudibility due to hearing loss, and distortion (if necessary); *c*) calculating the SNR for each band; *d*) applying a weighting to each band (because different parts of the speech spectrum carry more information than others); and *e*) summing the results. The basic process, that is, where no adjustments are made at *b*), can be summarized by the equation:

$$\mathrm{SII} = \sum_{i=1}^{n} I_i A_i \tag{2.6}$$

where n is the number of bands; I_i is the importance of each of those bands; and A_i is the proportion of speech cues audible in each band.

Unfortunately, as a predictor, the SII is not perfect and, whilst there is some correlation, there is no direct mapping between the SRT scores of normal-hearing listeners and the SII. This is partly because different intelligibility tests produce different SRTs, but also as a result of deficiencies in the SII itself. For example, although it is known that the SII is good at predicting intelligibility in stationary noise, it is less successful at predicting the SRT in quiet and even less successful at predicting it in fluctuating noise [Rhebergen and Versfeld, 2005], although proposed extensions to the SII have been shown to be more accurate predictors in fluctuating noise [Meyer et al., 2007].

An alternative to the SII is the Speech Transmission Index (STI) [Steeneken and Houtgast, 1980]. Although, like the SII, it is based on the AI, it has developed in

Table 2.1: *Comparison of speech intelligibility tests*

Name	Stimuli	Noise	Administration	Analysis	Time	Pros	Cons	Source
CVC	3 × 22 CVC nonsense syllables said after introductory sentence	Quiet	Say the syllable heard	% correct	N/K	No context; can isolate part phoneme	List equivalency. Training required. Not real speech	Fletcher and Steinberg [1930]
Rhyme Test	50 sets of 5 monosyllabic, CVC, rhyming words	Quiet or random	Write the letter heard to complete word	% correct	N/K	No context. Can be used in a variety of ways		Fairbanks [1958]
Modified Rhyme Test	50 closed sets of 6 meaningful CVC words differing in initial or final consonant with or without carrier sentence	Quiet or white noise	Choose from 6 alternatives	% correct or no. of confusions	≈25 min	Reliable. Fast. Commonly used	May overestimate. Risk of ceiling effect. Little diagnostic value	House et al. [1965]
Diagnostic Rhyme Test	96 pairs of meaningful CVC words differing in initial consonant without carrier sentence	Quiet	Choose 1 of 2 alternatives	% correct & diagnostic scores	≈15 min	No context. Reveals errors in discrimination of initial consonants. Guessing corrected for	May overestimate. Risk of ceiling effect. Restricted coverage	Voiers [1983]
Harvard/ PB words	20 phonemically-balanced sets of 50 common English words in carrier sentence	Quiet	Open or closed. Write word heard or multiple choice	% correct	≈30 min	Words have prosody	Training required for speakers. Learning effect	Egan [1948]
Harvard sentences	Fixed set of 100 phonetically balanced sentences	Quiet	Repeat sentence	% correct of 5 keywords	≈30 min	Easy to administer & score	Ceiling & learning effects. Restricted structure	Egan [1948]
SPIN	1 of 8 sets of 25 sentences	12-talker babble	Repeat last word of sentence	% correct words	N/K	Controls for semantic content	Whole sentence scoring	Kalikow et al. [1977]
HINT	25 sets of 10 sentences	Spectrally matched	Repeat sentence	% correct words from 5th sentence	<2 min	Very fast		Nilsson et al. [1994]
Haskins/ SNST	Fixed set of 100 semantically unpredictable sentences	Quiet	Repeat sentence	% correct sentences or keywords	≈30 min	No context	Limited in generalizability	Nye and Gaitenby [1974]
SUS	10 sentences in each of 5 different structures	Quiet	Repeat sentence	Correct/incorrect for whole sentence	≈15 min	No context. Different syntactic structures. Multiple languages	May be confusing. Long sentences. Not realistic	Benoît et al. [1996]
Matrix	Sentences in the form Name verb quantity adjective noun. Words in each slot chosen at random from list of 10	Various	Repeat sentence	Depends on test	≈15 min	Used in audiology. Flexible presentation. Multiple languages	Restricted sentence structure	Wagener [2009]
BKB-SIN	1 of 18 list-pairs of 10 BKB sentences	4-talker babble	Repeat sentence	Derive 50 % mark from % keywords correct	≈1.5 min	Low literacy requirement. Used for NH and HI	Semantic content	Niquette et al. [2003]
QUICKSIN	1 of 12 lists of 6 IEEE sentences (& 6 extra lists)	4-talker babble	Repeat sentence	Derive 50 % mark from keywords	≈1 min	Very fast	Female speaker	Killion et al. [2004]

a slightly different direction. Rather than inspecting and analysing the speech and noise signals directly, it treats the channel they are transmitted through as a 'black box' and applies a specially modulated test signal with speech-like properties to it. The reductions in the modulation caused by the channel are used as the measure of intelligibility or STI. Equation (2.7) shows how the calculation of the STI is actually quite similar to that of the SII in that both calculate an apparent SNR across a number of frequency bands, weight them, and sum them to arrive at an index between zero and unity. The STI uses seven bands, hence the maximum value of i, a weighting for each (w_i), and an apparent SNR ($\mathrm{SNR}_{\mathrm{app}_{i,j}}$) (30 is the assumed dynamic range of speech in dB and 15 is the limit of the apparent SNR range in dB).

$$\mathrm{STI} = \sum_{i=1}^{7} w_i \cdot \frac{\left(\frac{1}{14}\sum_{j=1}^{14} \mathrm{SNR}_{\mathrm{app}_{i,j}}\right) + 15}{30} \tag{2.7}$$

An experimental comparison of the SII and STI and a variation on the STI, the Rapid Speech Transmission Index (RASTI) found very little difference between them in terms of performance and, indeed, calculation methods [Larm and Hongisto, 2006]. The value of the two methods is that they are objective, require no listening tests, and can be implemented in hardware to produce devices that can be used to predict the intelligibility of a room or auditorium. Another common use is in predicting the efficacy of signal processing techniques in hearing aids. A promising area of research is in using the methods to adjust synthetic speech (or at least modify the choice of units) to make it more intelligible in noise [Cernak, 2006].

2.4 Synthetic speech intelligibility evaluation

Intelligibility tests have been used as a measurement for engineering purposes for a very long time, having seen considerable development during the early years of the telephone [Campbell, 1910]. The range of tests, and the manner in which they are employed in evaluating synthetic speech, are generally very similar to those set out in Section 2.2 with the obvious difference that many tests will include a comparison with natural speech. Again, the criteria for choosing an intelligibility test for synthetic speech such as those promulgated by Loizou [2007]—good coverage of the major speech phonemes, equal difficulty of test lists, and controlling for the effect of contextual information in the test material—are the same as one would use when testing natural speech. Within the field of speech synthesis, intelligibility tests may be used for a variety of reasons, perhaps the foremost being to compare TTS systems and

their components. Other reasons include: comparing synthetic speech directly with natural speech and assessing the intelligibility of synthetic speech to specific groups of listeners, such as those who are older, younger, non-native, or harder of hearing.

2.4.1 Comparing text-to-speech systems

As we saw in Section 2.2, intelligibility may be measured at various granularities of speech, from phoneme to whole sentence depending on need. When developing a TTS system, it is generally agreed [Benoît and Pols, 1992; Benoît et al., 1996; Taylor, 2009] that the choice of intelligibility measure should be related to the level of the TTS system being tested, such that unit testing might be based on phone- or word-level tests and system testing on sentence-level or comprehension tests [Taylor, 2009]. The tests described in Section 2.2 can be used to measure the intelligibility of synthetic speech with, perhaps, syllable-level tests being of particular benefit when testing a particular component of a TTS system for its production of, say, consonants, or features such as, voicing and nasality [Taylor, 2009] just as sentence-level tests may be more useful in assessing the sentence accenting and overall prosody of a system [Benoît et al., 1996].

The syllable-level tests that have commonly been used for testing particular components are the MRT and DRT. The MRT was criticized by Nye and Gaitenby [1974] as not being particularly useful for correcting synthesis errors although it was useful in identifying poorly synthesized phones. The reasons they gave were that the closed-response format forced listeners into inappropriate choices and that it did not provide full enough coverage of the language. The DRT may be useful for determining where errors in synthesis are occurring.

When it comes to comparing different systems, a systematic evaluation of many research, and some commercial, TTS systems has taken place every year since 2005 in the *Blizzard Challenge* [Black and Tokuda, 2005; Bennett and Black, 2006; Fraser and King, 2007; Karaiskos et al., 2008; King and Karaiskos, 2009, 2010, 2011, 2012]. Originally conceived by Black and Tokuda [2005], the first year saw six systems [Bennett, 2005] being put through their paces, with a maximum of nineteen seen in 2008 and 2009 [Karaiskos et al., 2008; King and Karaiskos, 2009]. Although the specific testing regime varies slightly from year to year, the challenge compares systems on naturalness, intelligibility, and (more recently) speaker similarity. Mean opinion score (MOS) tests are used for naturalness and speaker similarity, whilst forms of the MRT and/or SUS test are used for intelligibility. (The MRT-like test was dropped from 2007 onwards [Fraser and King, 2007].)

The first recorded use of SUSS for synthetic speech testing was by Nye and

Gaitenby [1974] who, as a result of their dissatisfaction with the MRT, devised the Syntactically Normal Sentence Test (SNST). More commonly known as Haskins sentences, after the lab in which they were developed, they comprise a closed set of 100 sentences. Having only one syntactic structure means that they are limited in generalizability and their ability to account for prosody as well as being susceptible to learning effects. However, the test is easy to conduct and tests a wide range of confusions.

Using Haskins sentences as a basis, the SUS test was developed by Benoît et al. [1996] with the twin aims of providing a test that could be used across a number of European languages and that covered a wider range of prosodic patterns and syntactic categories than previous tests. Implicit in the development of the test was the aim of controlling for the effect of context and semantic information and, therefore, ceiling effects that occur when testing highly intelligible TTS systems. (Benoît et al. [1996] stated that modern TTS systems could attain perfect intelligibility for simple and meaningful sentences and, although perhaps a little premature at the time of their writing, this is certainly the case today [King and Karaiskos, 2011].) As proposed, the SUS test truly is a sentence-level test, since the scoring is a binary correct/incorrect for the *whole* sentence. This, in turn, results in an intelligibility score noticeably lower than other (even sentence-level) tests; however, there is a correlation between them.

It should be noted that only the SUSs used in the 2009 *Blizzard Challenge* were generated using the original specification [King and Karaiskos, 2009]; those in previous years having been generated with the format 'Determiner adjective noun verb determiner adjective noun', such as, 'The unsure steaks overcome the zippy rudder' [Black and Tokuda, 2005] and scoring in all challenges has been on word error rate (WER) rather than whole sentences correct.

In addition to the obvious advantage that SUSs provide no contextual information to the listener, they use an open-response format such that an almost limitless number of sentences can be constructed, which means that they scale particularly well for use in large-scale comparison tests, such as the *Blizzard Challenge*. Furthermore, five different syntactic structures were devised to mitigate against learning effects, which has the added benefit that the test is very useful for differentiating between different prosody implementations or different speakers in the same TTS system.

However, it has been pointed out [Jurafsky and Martin, 2008; Taylor, 2009] that real applications tend to generate sentences that *are* highly predictable and so tuning TTS systems for SUSs may be counter-productive. There is also the possibility that some TTS systems may be better at producing what are essentially out-of-domain sentences than others.

It could be argued that speech does not have to be one hundred per cent intelligible

for the message to be fully received by the listener. For example, in the case of reminders, if a person hears, 'Take … amoxycillin', the fact that the word 'your' was not heard may make no difference to the comprehension of the message. In such circumstances, comprehension tests may be of value.

Pisoni [1987] was probably the first to use comprehension tests for the evaluation of synthetic speech at the end of the development of the *MITalk* system. For this he used standard adult reading comprehension tests and found that—for the second half of the tests—comprehension levels were equivalent to those achieved by reading the same passages of text. Subsequent studies of comprehension, summarized by Winters and Pisoni [2006], suggest that, whilst comprehension levels of the highly-intelligible *DECtalk* synthetic speech match those of natural speech, comprehension takes longer to achieve, thus suggesting that more cognitive resources are required.

More recently, Lines and Hone have used a form of comprehension test whilst trialling their interactive domestic alarm system [Lines and Hone, 2003b,a]. Their stimuli consisted of requests to move one of five shapes, in one of five colours to one of four rooms in their own home [Lines and Hone, 2003b]. It is difficult to draw conclusions on the efficacy of their method since only four people took part and it seems their performance was at ceiling level.

The motivation for comparing TTS systems, particularly on the scale of the annual *Blizzard Challenge*, is that it is the ultimate goal of speech scientists to make synthetic speech indistinguishable from natural speech. The obvious way to test for this is to compare synthetic and natural speech directly.

2.4.2 Comparing directly with natural speech

The review of synthetic speech perception and comprehension by Winters and Pisoni [2006] cites a number of studies undertaken between 1973 and 1993 that consistently show how synthetic speech is less intelligible than natural speech, at least at the segmental level using either open- or closed-response MRT tests (i.e., without any context). Although sentence-level intelligibility (i.e., with context) for the best TTS system at the time, *DECtalk*, was found to be statistically equivalent to natural speech [Mirenda and Beukelman, 1987]. The *Blizzard Challenge* of 2008 to 2012 [Karaiskos et al., 2008; King and Karaiskos, 2009, 2010, 2011, 2012] used SUS (the MRT-type test having been dropped in 2007), which test intelligibility without context. In the 2008 challenge, systems *T* and *V* were statistically indistinguishable from natural speech and six other systems were not significantly different from them. In 2009, system *S* was as intelligible as natural speech [King and Karaiskos, 2009] as were systems *J* and *R* in 2010 [King and Karaiskos, 2010] and system *C* in 2011 (2012

did not include a comparison with natural speech). So it seems that intelligibility has improved somewhat in the intervening years.

Many of the studies directly comparing synthetic and natural speech—and even those comparing different TTS systems—have used young, normal-hearing, native speakers of the language as listeners leading to some concerns about the generalizability of the results to the wider population and, particularly, to those falling into specific listener groups.

2.4.3 The intelligibility of synthetic speech to specific groups

Groups that can be identified that might have more difficulty in perceiving synthetic speech are: naive listeners; non-native listeners; younger listeners (i.e., children); those with a hearing impairment; and older adults.

Most people will have experienced having to 'tune in' to the speech of someone with a strong regional accent or speech impairment. A similar process occurs, at least subconsciously, with any speech we listen to and especially when that speech is synthetic. It stands to reason, then, that a regular listener to a synthetic voice will perform better on an intelligibility test than a naive listener. In order to control for this in experiments, it is common practice to screen out regular listeners or to provide a period of familiarization to all participants.

Greene [1986] first compared native and non-native listeners' performance on a MRT and a dictation task and showed that non-natives performed worse than natives and that their performance was related to their proficiency in English. A similar study, but this time using semantically anomalous sentences [Mack, 1988] found a similar disparity between the two groups and suggested that the non-natives tended to listen out for the wrong cues, perhaps influenced by their other (in this case, German) language. Later studies in noise [Reynolds et al., 1996; Alamsaputra et al., 2006] concurred with the earlier studies and showed that the added noise affected non-natives more than natives. The lower performance of non-natives in both quiet and noise means that this aspect needs to be controlled for in perceptual experiments.

In a similar vein, children's language skills are thought to be the cause of their consistently poorer performance on intelligibility tests in both quiet and noise. Since the intelligibility of synthetic speech to children forms no part of the proposed work, the interested reader is directed to Drager and Reichle [2010] for a comprehensive review of the literature relating to this area of research.

People with hearing impairments have been shown to find synthetic speech less intelligible than those with normal hearing, both in quiet [Kangas and Allen, 1990; Nixon et al., 1990; Humes et al., 1991] and in noise [Nixon et al., 1990], so hearing

loss was accounted for in the experiments we ran, but no systematic investigation of the effects of different types of loss was undertaken. A number of participants did have some hearing loss, not least because it forms part of the normal ageing process.

Presbycusis is the term that describes the loss of hearing that gradually occurs in most individuals as they grow older [Katz et al., 2009]. Corso [1959] found it to be a (usually) high-frequency sensorineural loss that progressively spreads down to the lower frequencies; affects men earlier—and more—than women; is more diverse in its presentation in men; and tends to affect both ears about equally. More recent studies concur with these early findings, although it has been shown that women suffer more low-frequency loss [Jerger et al., 1993]. For a detailed review of work on age-related hearing loss, see Chisholm et al. [2003]. Traditionally, hearing ability has been measured using pure-tone audiometry. As the name suggests, pure-tone audiometry measures the lowest intensity at which various pure tones can be perceived, but usually only up to 8 kHz [British Society of Audiology, 2004; Katz et al., 2009]. Because age-related hearing loss begins at higher frequencies, it is possible for a person to 'pass' a pure-tone audiometry test and still have difficulty hearing speech.

A confounding factor in measuring intelligibility in older adults is the possibility of cognitive decline. Whilst it is self-evident that those with diagnosed dementia could be expected to have difficulty in perceiving natural and human speech, 'normal' cognitive ageing does not appear to have an effect. Simulating age-related hearing loss, by presenting MRT stimuli in spectrally-shaped noise to younger listeners, Humes et al. [1991] demonstrated that older age had no discernible effect on intelligibility. However, Roring et al. [2007] cast doubt on the applicability of diphone-based synthetic speech commonly used for AAC devices. Studies using stimuli generated by unit selection [Lines and Hone, 2003b; Wolters et al., 2007] produced similar findings and found that older adults' comprehension was not significantly different from younger persons', given sufficient context. The importance of context had earlier been demonstrated by Drager and Reichle [2001], who also showed that the benefit gained by older people was not significantly different from that gained by younger people.

2.5 Synthetic speech intelligibility in noise evaluation

Background noises may be used in the evaluation of synthetic speech just as they are for natural speech (see Section 2.2.4). The type and level of noise used will depend on the use to which it is being put. The uses will be the same as when evaluating synthetic speech in quiet with the addition of testing the effect on intelligibility of the noise itself (usually in a particular domain); however, only those pertinent to this

thesis are detailed in the following sections.

2.5.1 Comparing text-to-speech systems in noise

The *Blizzard Challenge* is the largest known comparison of systems currently carried out and, as previously mentioned, a number of the systems are approaching—or have already achieved—the level of intelligibility of natural speech. This means that there is a real risk of ceiling effects occurring, so, to facilitate finer discrimination between systems, the listening test needs to be made harder, which can be achieved by introducing a background noise. The *Blizzard Challenge* 2010 introduced a speech-in-noise test for the first time.

Nearly all previous studies of synthetic speech intelligibility in noise have focused on formant synthesis and diphone concatenation systems. Since modern speech synthesis methods such as unit selection [Hunt and Black, 1996] and HMM-based synthesis engines [Zen et al., 2009] yield speech that is more intelligible than formant and diphone synthesis, we would also expect them to be more intelligible in background noise. This assumption is supported by the results of Lancaster et al. [2004], who found that an AT&T unit selection voice was more intelligible than *DECtalk*'s *Perfect Paul* in cockpit noise at SNRs of −5, −8, and −11 dB.

As far as can be determined, there had been no *systematic* investigation of the intelligibility of HMM systems, and only one [Lancaster et al., 2004] of unit selection, under different SNRs. While van Leeuwen and van Balken [2005] seem to imply that they tested unit selection systems in their experiment, they do not specify the synthesis methodology used for any of their systems.

Venkatagiri [2003] tested four of the most up-to-date products at the time, one formant-based, one diphone-based, one half-phone-based, and a hybrid concatenative/formant coding system. He used the MRT presented to normal-hearing listeners in a 20-person multi-talker babble at SNRs of 0 and 5 dB and found that no system resisted noise better than the others, but all of them fared better at 5 dB than 0 dB. The hybrid system had the worst intelligibility of all.

Venkatagiri [2003] also compared the speech of the four TTS systems with natural speech presented in a 'comparable' manner. Even the best TTS systems were found to be 22 % less intelligible than natural speech, which, he concluded, made them unlikely to be suitable for unrestricted output.

2.5.2 Comparing directly with natural speech in noise

Very few other studies seem to have been published that directly compare the perception of natural and synthetic speech in the presence of noise of normal hearing adults.

Pisoni and Koen [1982] found that the intelligibility of synthetic speech suffered more than natural speech over a range of SNRs, using both an open- and closed-set MRT. Using an open-response version of the MRT and a twelve-talker babble as background noise at levels of 0, 15, and 25 dB, Koul and Allen [1993] found that the highly intelligible *DECtalk Paul* and *Betty* voices suffered more than natural speech from the effects of the interfering noise; although, interestingly, they found no significant difference between the male and female voices.

2.5.3 Measuring the effects of noise in various domains

Whilst some work has been done on assessing the effect of background noise on synthetic speech, it has often focused on systems used in specialized situations, such as over the telephone [Möller, 2004; Langner and Black, 2005], whilst driving [Morrison and Casali, 1997], or in aircraft cockpits [van Leeuwen and van Balken, 2005]. The telephone domain is highly complex, because noise can be present at the speaker's end and at the listener's end and it can be introduced during transmission. For an overview of the specialized assessment procedures required, see Möller [2004]. For contexts such as driving a car or flying an aeroplane, a masking noise can be used that represents typical spectral and temporal characteristics of background noise in these situations.

Although some studies of the intelligibility of synthetic speech have been undertaken in the home environment [Lines and Hone, 2003b,a], none appears to have evaluated the effect of the noises found there.

2.5.4 Measuring the effects on older listeners

Langner and Black [2004, 2005] compared the effect of noise on older adults' perception of synthetic speech as part of their work on creating a database of 'speech-in-noise'. (Here the term 'speech-in-noise' refers to the fact that the speaker was asked to speak in the presence of noise, so that the speech would be modified to make it more intelligible in noise. They felt the term 'Lombard speech' [Lombard, 1911] inappropriate, as the level of background noise was small and they did not address extremes of speech.) The aim was to record a voice talent speaking in quiet and with noise being played through headphones, which, they believed, would allow them to extract the differences between the two sets of speech so that any voice could be made more robust in noise via 'style conversion'. Unfortunately, 'the style-converted synthetic speech was nearly universally harder to understand than the original plain speech' [Langner and Black, 2005, p. 396]. However, their reports do suggest that older adults have much more difficulty than younger adults in perceiving synthetic

speech at the SNR of −3.2 dB that they used [Langner and Black, 2004, 2005].

McCarty and Surprenant [2006] set out specifically to test the intelligibility of AT&T's *Natural Voices* synthesizer to older adults at various SNRs as compared with natural speech using the SPIN test [Bilger et al., 1984]. An abstract produced by them suggests that older adults found synthetic speech universally harder to understand *and remember*. Unfortunately, their findings have not been written up as a complete paper, leaving a number of intriguing questions, such as, whether they compared older adults with younger adults and what controls they had put in place.

2.6 Summary

The flexibility of synthetic speech in its ability to generate any valid utterance and to be configured for optimal output in various conditions makes it the preferred choice over canned speech for many applications. Its importance as a means of output is manifest in its continued development from formant and diphone systems through to the current state-of-the-art systems based on unit selection or HMMs.

Evaluation of the intelligibility of synthesis systems has followed the model adopted for the assessment of natural speech, that is, measuring at the syllable, word, or sentence level, often using noise as a discriminating factor. Although many of the materials and methodologies for the assessment of natural speech have been brought to bear in the assessment of synthetic speech, much research has been orthogonal, with studies focusing on either natural or synthetic speech. A similar dichotomy is evident between research in audiology and synthetic speech intelligibility.

A close inspection of the research literature reveals, not only a separation of synthetic speech research from other domains, but also from the real world. Many of the stimuli and masking noises do not occur in real-life situations and even experimental participants tend to come from a very narrow section of the population. Our intention in the following chapters was to develop and implement an experimental methodology to bring together these various threads into a cohesive approach to assessing synthetic speech.

CHAPTER 3

EXPERIMENTAL METHODOLOGY

THIS work was undertaken in the context of the provision of spoken reminders in the home environment and having identified background noise and reverberation as the biggest hindrances to their effective use. We determined to carry out a series of experiments to establish the effects of each on current synthesis systems.

We, therefore, needed to identify techniques that could be used in all our experiments that were ecologically valid—that is, approximated the real-world—whilst maintaining scientific rigour. We refer to these as our 'experimental methodology'.

The literature review presented in Chapter 2 reveals little recent work in evaluating the intelligibility of synthetic speech in noise. The two studies that had been published and used more recent synthesis systems had used aircraft engine noise at signal-to-noise ratios (SNRs) of −5, −8, and −11 dB with no demographic data on participants provided [Lancaster et al., 2004]; or a noise 'spectrally equivalent to cockpit noise' and a speech reception threshold (SRT) methodology with participants who were only described as 'Dutch' [van Leeuwen and van Balken, 2005, p. 2].

The largest study of synthetic speech intelligibility in quiet, the *Blizzard Challenge*, had been run annually using mainly undergraduate students and speech synthesis researchers as listeners; and, since 2007, semantically unpredictable sentences (SUSS) as stimuli [Fraser and King, 2007].

None of the techniques used in operationalizing the studies reviewed fully met our needs, so we investigated alternatives. The remainder of this chapter documents our assessment of the suitability—for use in the remainder of the thesis—of those identified: Amazon Mechanical Turk (AMT); Matrix sentences; and one of the noise files created by the International Collegium of Rehabilitative Audiology (ICRA).

3.1 Amazon Mechanical Turk

Assessing the intelligibility of synthetic speech has traditionally been carried out by inviting a number of participants to listen to speech stimuli in specially designed booths that allow for the control of confounding variables, such as quality of listening equipment and background noise, thus allowing for a definitive assessment of a system's intelligibility. However, more often than not, the absolute performance of a system is less important that its performance relative to another and the cost (in money and time) of maintaining a bespoke laboratory and recruiting participants becomes hard to justify.

Moreover, the nature of experiments carried out in listening labs in universities is that they attract primarily undergraduate students, who, by definition, tend to have a very specific age range and education, and who may become 'serial testers' and, therefore, not the naive listeners originally intended. Even without the preponderance of students, labs limit the number of participants that can be recruited, which in turn means that that small, but significant, effects may not be found [Bunnell and Lilley, 2007].

In recognition of these concerns, many researchers, including the organizers of the first *Blizzard Challenge* [Black and Tokuda, 2005], began conducting listening tests over the Internet. Although this removes the time and expense of running lab sessions and travel expenses for participants, the problems of finding and paying participants remain and new problems, such as the possibility of the same person participating more than once, are introduced.

A possible solution for these remaining problems is the marketplace for crowdsourcing developed by Amazon and known as Amazon Mechanical Turk[1]. Crowdsourcing may be defined as, 'The practice of obtaining information or services by soliciting input from a large number of people, typically via the Internet and often without offering compensation.' [OED Online, 2013], although Amazon places a specific emphasis on tasks that require human intelligence and that cannot currently be completed by computers.

Although the term may be relatively new, the practice is not. As already stated, synthetic speech has traditionally been assessed by soliciting large numbers of people to listen to it and provide input on its intelligibility or naturalness; often using the Internet as a recruitment tool or, as in the case of the *Blizzard Challenge*, as a means of conducting listening tests remotely.

Amazon has commercialized crowdsourcing by implementing an Internet-based system that simplifies and streamlines the process of advertising for, recruiting, and

[1] `http://www.mturk.com`

paying participants such that AMT has become the pre-eminent of such services.

As a result, crowdsourcing is increasingly used to create rich speech and language data sets [Callison-Burch and Dredze, 2010]. Instead of relying on highly skilled annotators and transcribers, researchers ask anonymous Internet users to contribute annotations [Snow et al., 2008], transcriptions [Novotney and Callison-Burch, 2010; Marge et al., 2010], and ratings [Kittur et al., 2008] for as little as a cent per task.

Of course, AMT is not without problems of its own with rumours of participants 'gaming' the system—that is, giving answers that sped them through completion of the task rather than the most accurate—being particularly prevalent. Furthermore, the difficulties associated with any system for running experiments at a distance, such as conducting hearing screens, remain. Ultimately, there will invariably be a trade-off between running more experiments with more participants and fewer, smaller experiments with more control.

3.2 Matrix sentences

Comparison of intelligibility between complete text-to-speech (TTS) systems is generally undertaken using sentence-level tests, of which the SUS test [Benoît et al., 1996] has become the de facto standard. The test, as originally defined, allowed for the use of five syntactic structures. In perhaps the most well-known test, the *Blizzard Challenge*, SUSs of the form 'Determiner adjective noun verb determiner adjective noun' are generally used and scored on individual word error rate (WER) with each word carrying an equal weight. (Except in the case of the *Blizzard Challenge* 2009, when a mixture of the original SUS structures were used, but scored on WER rather than sentences-correct [King and Karaiskos, 2009].)

Benoît et al. [1996] developed SUSs to control for context and semantic information and, therefore, the ceiling effects that can occur when testing highly intelligible TTS systems. However, as with all sentence-level tests they may suffer from memory effects, since the respondent is expected to remember the whole sentence—rather than a single word—before responding. A further problem, in the context of our research, was that the lack of meaning in the sentences undermined their ecological validity.

Ideally, we needed sentences that lacked context and, therefore, predictability, but had meaning whilst having a length similar to that of a typical spoken reminder. More generally, we were interested in finding stimuli that did not suffer from the shortcomings identified in Chapter 1, namely the possibility that they tended to tune TTS systems towards producing non-realistic utterances, or that they were biased towards systems that were inherently better at creating such utterances, for example, favouring the HMM-based Speech Synthesis System (HTS) over *Festival* unit-selection

(F-USEL) [Yamagishi et al., 2009].

Matrix sentences [Wagener, 2009], discussed in Chapter 2, appeared to meet all the above criteria, with the additional benefit that they had value outside synthetic speech intelligibility testing, since they had been developed as an audiological test and, thus, raised the possibility of equating hearing ability with synthetic speech intelligibility. They have advantages over SUSS in that they are less difficult, have meaning, are typically shorter, and are, therefore, well suited to reminders. In experimental conditions, they have the additional advantage that fixed-format sentences from a limited set of words are far less likely to suffer from spurious insertions, that is, participants asserting that they had heard a word that was not in the stimulus.

One of the drawbacks of using short meaningful sentences is that they can lead to ceiling effects, but this was accounted for in the choice of words used in the development of the sentence materials. Moreover, as we envisaged conducting all of our testing in the presence of noise, ceiling effects were not as likely to be encountered.

A disadvantage of using fixed-format sentences with a limited number of words is that participants need to be trained in their use in order to obviate any possible learning effect. This can be accounted for by always providing training at the start of the experiment and using statistics to measure the extent of any learning effect that occurred.

3.3 Background noises

Our intention for our experiments involving noise was that we would use noises that were likely to occur in the home environment. However, our literature review had revealed that speech intelligibility in noise is typically assessed using highly standardized, generic noises such as white noise, pink noise, and multi-talker babble. Those studies that had used realistic background noises often focused on those from specific scenarios, such as using the telephone, driving, or flying an aeroplane, rather than noises likely to be heard in the typical home.

Moreover, many of the noises are stationary, that is, they do not vary over time, whereas speech and other man-made noises tend to be non-stationary, or fluctuating. The SNR at which fifty per cent of speech is intelligible (the SRT) [Brand, 2009] varies depending on the type of background noise. The SRT in fluctuating noise occurs at a SNR of about -12 dB as opposed to -4.5 dB in stationary noise [Rhebergen and Versfeld, 2005], which obviously has implications for our research.

We intended to use AMT to establish what noises were, in fact, common in our participants' homes. In the meantime, we wanted to investigate the intelligibility of unit-selection- and hidden Markov model (HMM)-based synthetic speech in everyday

and home environments, using a fluctuating noise, as these are more realistic than stationary noises [Rhebergen and Versfeld, 2005], so we hypothesized that the most common background noise encountered in these environments is human speech, either directly or indirectly via the television or radio. We, therefore, initially chose a noise from the ICRA suite of noises [Dreschler et al., 2001].

The noises were developed for the International Collegium of Rehabilitative Audiology by the Hearing Aid Clinical Test Environment Standardization (HACTES) working group in order to establish a collection of background noise signals that could be used in testing hearing aids and other instruments. They are widely used in audiology and audiological research, including by the HearCom project in the creation of the Matrix sentences [Wagener et al., 2007]. The noise signals are composed of real speech modified with defined spectral and temporal characteristics to provide real-life speech and babble noise.

The number of speakers in the background noise has an effect on the intelligibility of the target speech, so we chose a noise that emulates a single speaker on the assumption that this is the most likely scenario. Also, since it is known that understanding speech in speech-shaped noise is more difficult if the noise matches the gender of the speaker's voice [Drullman and Bronkhorst, 2000], we always selected a noise spectrally matched to the gender of the speaker used to generate the speech part of the stimuli.

As Christensen et al. [2010] point out, humans alter their speech in the presence of noise, so adding a noise to the same speech stimulus at various SNRs may result in a mixed signal that would never be encountered in real life; however, since speech synthesis systems do not currently make any adjustments to their output in response to background noise, we created stimuli by simply adding background noise to the speech.

3.4 Methodology

We evaluated synthetic speech, either alone or with natural speech, using human participants in listening experiments, using a mix of AMT and the lab. Statistical analyses were carried out using mixed model analysis [Gelman and Hill, 2006; Baayen, 2008], which is explained in Section 3.4.2.

3.4.1 Overview of experiments

We carried out four initial experiments. The first established a baseline for future experiments; the second compared AMT with the lab; the third compared Matrix

Table 3.1: *Overview of initial experimental-methodology experiments*

Experiment	Purpose	Systems	Stimuli	Noise	SNRs	Design	Participants	Sec.
Baseline	How intelligible are current TTS systems?	*Nick*: HTS v. F-USEL	50 Matrix sentences	ICRA track 5	−10 dB to 10 dB in 5 dB steps	Within-subjects. Each heard both systems and all SNRs in 1 of 10 balanced lists of 50 sentences	Native UK English: 20 in perception labs	3.5
AMT	Can AMT replace the lab?	US *KAL*: Diphone v. HTS UK *Nick*: F-USEL v. HTS	50 SUSS	None	N/A	Between-subjects. Each heard the same 50 sentences in one system/voice	Native US English: 159 on AMT; 20 in a quiet room	3.6
Matrix	Can Matrix replace SUSS?	*Nick*: HTS v. F-USEL	50 Matrix sentences	None	N/A	Between-subjects. Each heard the same 50 sentences in one system/voice	Native US English: 59 on AMT	3.7
New voice and natural speech	Can AMT be used with another voice and natural speech?	*Roger*: HTS v. F-USEL v. NS	60 SUSS	None	N/A	Within-subjects. Each heard 1 of 3 balanced lists of 60 SUSS	Native US English: 58 on AMT Native UK English: 24 in perception labs	3.8

sentences with SUSS; and the fourth confirmed the utility of AMT with another voice and natural speech. Each experiment followed the same general format with the main changes being the format of the stimuli used as required to suit the purpose of the experiment. A summary of experiments is provided in Table 3.1.

No participant took part in more than one of the four experiments. Each experiment was delivered using a series of web pages, broken down into parts. The first part collected a standard set of demographic data for each participant; the second technical data; the third data about the participant's hearing; the fourth part presented each stimulus on a separate web page and collected the typed response; and the last collected supplementary data. The demographic and other data gathered are detailed in Table 3.2.

In accordance with Amazon's guidelines at the time, participants were permitted to answer, 'prefer not to say' (which was recorded as 'withheld') to the demographic questions, except place of birth and dialect. All items were self-reported. The same information was collected for all subsequent experiments, except where stated.

3.4.2 Mixed model analysis

We used R [R Development Core Team, 2009] for all descriptive and analytical statistics. We modelled the effect of speech synthesis system type on the number of errors participants made using generalized linear mixed models [Gelman and Hill,

Table 3.2: *Demographic and other data collected from experiment participants*

Part	Data item	Permitted responses
Demographics	gender	male; female; withheld
	age range	under 20 to over 80 in ten-year groups; withheld
	education	before high school; high school; some college; bachelor's degree; master's degree; doctorate; withheld
	occupation	retired; home maker; employed full-time; employed part-time; student full-time; other; withheld
	place of birth	United States; Canada; United Kingdom; Australia; New Zealand; Ireland; Other
	dialect	US or Canadian; UK; Irish; Australian; New Zealand; South African
Technical	computer scientist/engineer	yes; no
	work in speech technology	yes; no
	how often listen to synthetic speech?	at least once a week; at least a couple of times a year; rarely or never; I'm not sure
	headphone type	ear buds; in-ear; on-ear; full-ear
	headphone features	noise cancelling; sound isolating; none that I know of
Hearing	hearing aids	no hearing aids; left ear; right ear; both ears
	ten questions of the HHIA-S	yes; sometimes; no
	other hearing information	free-form text
Supplementary	environment	quiet all the time; quiet most of the time; equally quiet and noisy; noisy most of the time; noisy all the time
	kind of noise heard	N/A; radio/television; conversation; music; traffic; domestic appliance
	noise characteristics	N/A; constant; fluctuating; in short, isolated bursts
	web browser	Internet Explorer; Firefox; Mozilla; Opera; Safari; Chrome; Other
	current location	New England; Mid-Atlantic; East North Central; West North Central; South Atlantic; East South Central; West South Central; Mountain; Pacific
	experience of stimuli	usually understand all of the words; usually understand most of the words; words were very hard to understand; I have trouble typing
	any other comments	free-form text

2006; Baayen, 2008]. Mixed models are statistical models that account for both fixed and random effects.

The term 'fixed effects' refers to those presumed to be caused by the explanatory variables under investigation and are repeatable; whereas the term 'random effects' refers to those presumed to be caused by the latent variables introduced by the experiment design and are not repeatable.

For example, we might ask 20 participants to listen to 50 different sentences synthesized by two TTS systems and record what they heard, measuring the proportion of words they heard incorrectly as the WER. We would expect the effect of TTS on WER to be 'fixed' and to see similar overall WERs if we ran the same experiment with 20 new participants. On the other hand, we would not expect any individual participant in the second experiment to have exactly the same WER as any participant in the first; nor for any particular sentence to return the same WER as any other if we used 50 new sentences. That is to say, the WER of any particular participant or sentence is unpredictable, or 'random'.

In simple terms, mixed models are an extension of the traditional linear model, the formula for which is shown (in matrix terms) in Equation (3.1) where y is a vector of responses (in our example, WERs); X is the design matrix of explanatory variables; β is an unknown vector of fixed effects; and ϵ is an unknown vector of random errors.

$$y = X\beta + \epsilon \tag{3.1}$$

Mixed models explicitly model random effects by using a design matrix Z. Equation (3.2) shows the equivalent formula for a mixed model including the additional term, $Z\gamma$, where γ is an unknown vector of random effects and Z is its design matrix. Essentially, using our example, the individuality of participants and sentences has been accounted for and 'partitioned' in the new term rather than being allowed to spuriously affect the fixed effects or unnecessarily increase the number of errors. This means the model more accurately reflects the real world allowing it to provide better predictions.

$$y = X\beta + Z\gamma + \epsilon \tag{3.2}$$

A further advantage is that the errors in the error term (ϵ) no longer have to satisfy the requirement of being independent of one another. Mixed models confer further advantages over the alternatives, such as analysis of variance (ANOVA), in that they can cope well with missing or sparse data and automatically account for nested designs.

Models throughout this thesis were fitted using the *R* [R Development Core Team, 2009] package *lme4* [Bates and Maechler, 2009]. The reporting of models often includes reference to the Akaike information criterion (AIC) [Akaike, 1974],

a measure used to compare models. Although not an absolute test of a model, it provides a measure of its quality relative to other models by measuring the trade-off between how well the model fits the data and how complex it is. The lower the value it returns, the better the model. The significance of the difference between models can be established by means of the χ^2 statistic returned from an ANOVA comparison.

3.5 Baseline experiment

The review of the speech synthesis literature presented in Chapter 2 revealed that little work had been carried out on assessing the intelligibility of modern synthetic speech systems in the presence of a background noise and none on the effects of reverberation. This first experiment, therefore, was conceived as a baseline to establish the merit in further investigation and to trial the use of Matrix sentences and the ICRA noise as potential vehicles for carrying out future studies.

Our main concern was to establish a baseline set of data that illustrates how unit selection and HMM systems perform in a validated, fluctuating, background noise appropriate to home (and other) environments. Secondly, we wanted to see which system is better suited to providing speech in the presence of noise. Finally, we sought to identify SNRs that would be useful for future experiments.

3.5.1 Method

Since we wanted to establish a baseline for the intelligibility of TTS systems in background noise, we chose to use two of the baseline systems from the *Blizzard Challenge* 2009 [King and Karaiskos, 2009]; namely, F-USEL and the 2007 variant of HTS. We used the high quality RP-English voice *Nick*[2] created for the *Multisyn* unit-selection engine [Clark et al., 2007] for both systems, which means that only the synthesis method is varied, not the underlying speech data.

We used one of the ICRA noises described in Section 3.3, specifically track number 5 from the ICRA compact disc (CD). The track is 5 min 2.25 s in length and consists of two channels of 3-band speech-modulated noise. It is has a male-weighted idealized speech spectrum at normal effort and, therefore, closely matches the spectral and modulation characteristics of the natural speech of a male speaker.

We presented the stimuli at SNRs of −10, −5, 0, 5, and 10 dB. The lowest SNR is close to the SRT in fluctuating noise for human speech, around −12 dB [Rhebergen and Versfeld, 2005], and the highest SNR is close to a value where ceiling effects may occur.

[2] http://www.cstr.ed.ac.uk/projects/festival/onlinedemo.html

Fifty unique sentences, in ten batches of five, were generated at random from the word list given by Wagener [2009], ensuring that each word used occurred exactly five times. In our design, there is a total of ten conditions (2 systems × 5 SNRs). Using a Latin square design of order ten, we created ten lists of sentences, ensuring that each of the fifty sentences occurred in each of the ten conditions across our lists.

In order to ensure that both HTS and F-USEL used the same speaking rate, we standardized the durations of both sets of stimuli. The mean durations of both sets of speech files were calculated and the ratio of these means, 0.807 62, was used as the adjustment factor to regenerate the HTS files. We only changed the duration of HTS files because it allows for this to be done without unduly distorting the output (as would occur by simply applying a linear 'stretch'). The final mean sentence duration of F-USEL sentences was 1.88 seconds, with a standard deviation of 0.2 and the mean duration of the HTS sentences was 1.91 seconds, with a standard deviation of 0.1. The sound pressure level (SPL) of both sets of files was then adjusted to 65 dB to reduce the possible effect of differences in amplitude.

Noise was added to each of the speech files by taking a matching-length, random sample from the original noise file for each speech file at each of the five noise levels; thus giving a total of 500 speech-with-noise files. Samples were not taken from the first 350 ms and last 2 s of the ICRA noise, as these only contained silence. In standardizing the SPLs and in calculating the SNRs, we used the active speech level as determined by method B of the ITU-T P.56 standard [ITU, 1993] using the *MATLAB* implementation of Loizou [2007]. The active speech level is designed to match more closely what the human ear actually perceives.

A total of 20 volunteers (15 male and 5 female) from staff and students at the University of Edinburgh took part in the experiment. The spread of ages was: 1 under 20; 8 20 to 29; 8 30 to 39; and 3 40 to 49. Participants were asked to record the highest level of education they had completed: 3 had completed high school; 3 a bachelor's degree; 6 a master's degree; and 8 a doctorate. All participants were screened for normal hearing using the Hearing Handicap Inventory for Adults Screening Version (HHIA-S) [Newman et al., 1991], substituting *cinema* for *movies* and using British English spellings. All participants scored below ten, the threshold at which a hearing test would be recommended. The mean score was 1.1, with a minimum of 0 and a maximum of 8. All participants were native speakers of UK English.

Participants listened through Beyerdynamic DT770 PRO 250Ω headphones fed by a Focusrite Saffire LE external FireWire sound card on an Apple Mac Mini in an isolation booth at the Centre for Speech Technology Research (CSTR). Two participants were assigned to each list of sentences.

The composition of the Matrix sentences was explained to participants and they

were given six sentences without noise with which to practise. They were asked to adjust the volume of the headphones to a comfortable level at the start of the test and not to adjust it further. They were prevented from playing any file more than once. Instructions were given at the presentation of each file to type in what they thought they heard and to guess when unsure. One file for one respondent failed to play and the response was left unscored.

In order to be able to compare intelligibility without undertaking unnecessary error analysis, we measured participant performance using WER [Bunnell and Lilley, 2007]. This was calculated using the *Blizzard Challenge* data analysis script[3], which determines the total 'cost' of transforming the participant's response to the correct string as the sum of the number of insertions, substitutions, and deletions divided by the number of words in the correct string. For our analysis, we used the total number of errors, since the number of words was the same in each sentence. The resulting variable follows a Poisson-like distribution.

3.5.2 Results

Results are summarized in Figure 3.1. For a SNR of 0 dB and better, results mirrored what was known from comparative evaluations, such as the *Blizzard Challenge* [King and Karaiskos, 2009], when HMM-based systems consistently outperformed pure unit selection. For 5 and 10 dB, performance appears to be at ceiling, with next to no errors. At −5 dB, participants make an average of one error per sentence for both systems—the advantage of HTS over F-USEL disappears. At −10 dB, we even find F-USEL marginally outscoring HTS (mean number of errors for HTS: 2.21; mean number of errors for F-USEL: 2.01).

We examined the data in more detail using a generalized linear mixed model (GLMM) [Gelman and Hill, 2006]. Since the mean number of errors is 0.79, and the variance is 1.44, the response variable is overdispersed. The negative-binomial distribution is commonly used to model overdispersed count outcome variables and can be seen as a more generalized form of the Poisson distribution, since it has the same structure, but with an additional parameter to model the dispersion[4]. Figure 3.2 shows how the negative-binomial distribution fits our data better than the Poisson.

In order to evaluate our choice of the negative-binomial model, we carried out a likelihood ratio test to compare Poisson and negative-binomial generalized linear models of the data, which supported the choice ($\chi^2(1) = 6.69$, $p = 0.01$). An examination of the quantile–quantile (Q–Q) plot of the residuals of the generalized

[3] http://www.cstr.ed.ac.uk/projects/blizzard/tools.html

[4] We originally used the quasi-Poisson family in our model instead of the Poisson family [Gelman and Hill, 2006], but this has now been deprecated.

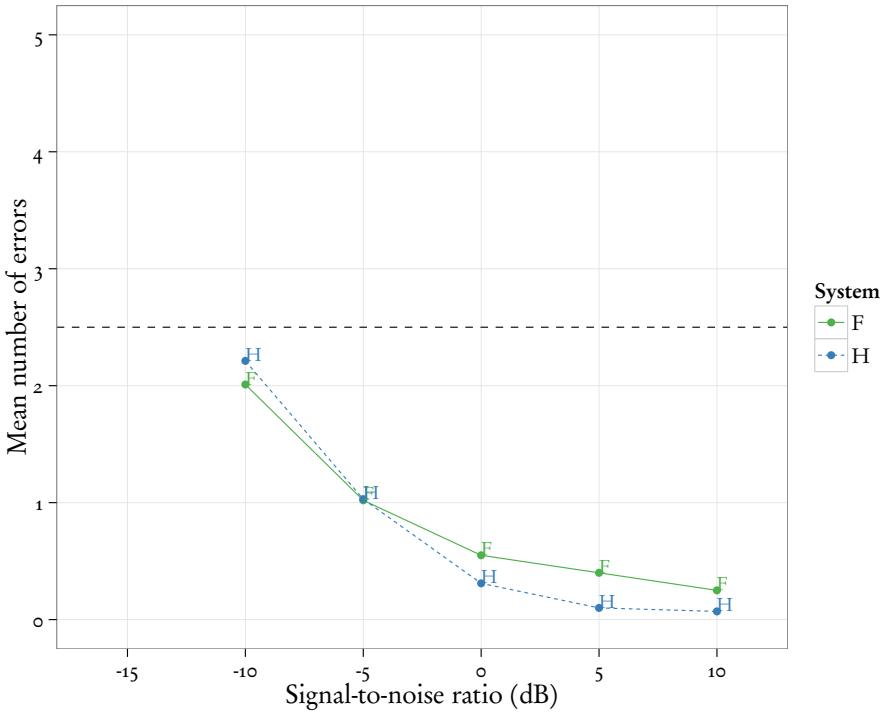

Figure 3.1: *Mean number of errors for F-USEL (F) and HTS (H) at all SNRs. Dashed horizontal line indicates error level at SRT*

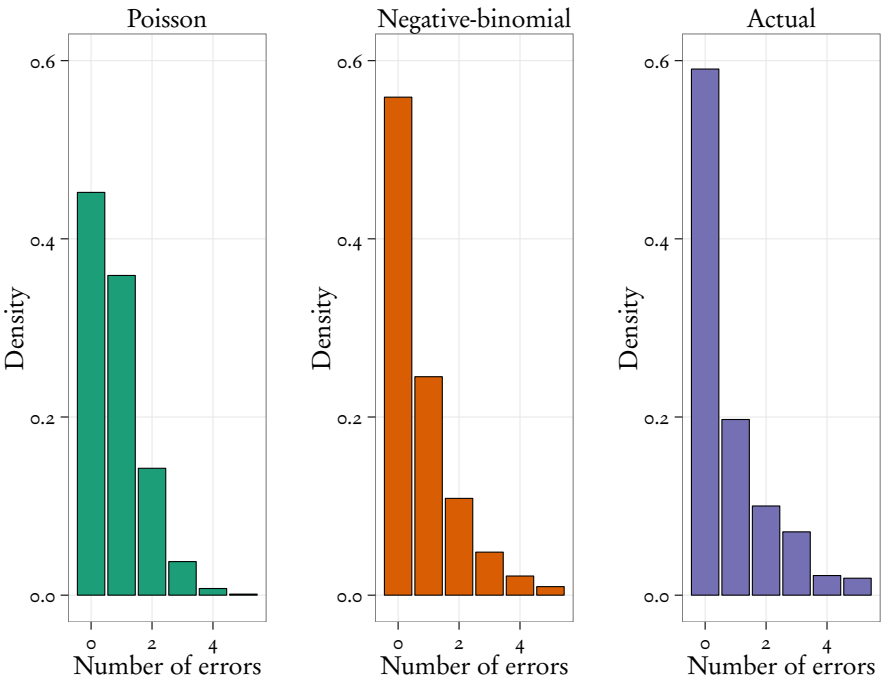

Figure 3.2: *Comparison of predicted Poisson and negative-binomial distributions of numbers of errors to the actual distribution*

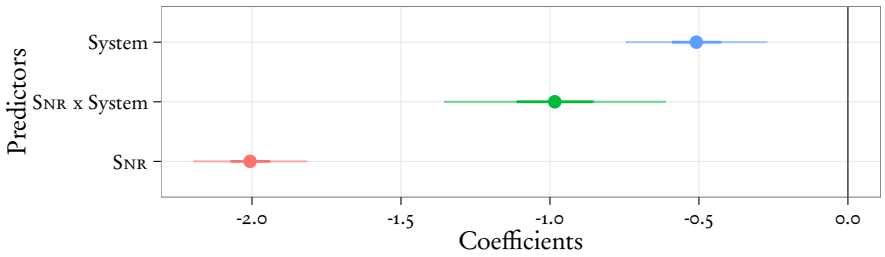

Figure 3.3: *Coefficients of individual-level predictors System, SNR × System, and SNR. Dots correspond to point estimates, lines to 95 % confidence intervals*

linear mixed model confirmed its fit to the data. The model was fitted using the *lme4* package [Bates and Maechler, 2009] with the following formula:

$$\text{Error} \sim \text{SNR} * \text{System} - 1 + (1 \mid \text{Participant}) + (1 \mid \text{Sentence}) \tag{3.3}$$

It consisted of two individual-level predictors, SNR and System, the interaction of both predictors, SNR × System, and two group-level predictors, the identifiers for participant and sentence. System is 1 if the system is HTS; 0 if it is F-USEL. We only varied the intercept for each of the group-level predictors. Sentences that are more difficult will contain more errors, and the higher error rate will be reflected in a higher sentence-level intercept. Similarly, some people are more likely to mishear words than others, and people who often get words wrong will have higher intercepts. Thus, our group-level predictors allow us to correct for overall sentence difficulty and overall performance differences between participants.

Following Gelman [2008], the individual-level predictors were rescaled to yield a mean of 0 and a standard deviation of 0.5, which makes it easier to compare the effect of the binary predictor, System, and the interval-sized predictor, SNR.

Figure 3.3 shows point estimates (dots) and confidence intervals (lines) for the coefficients of the individual-level effects. Effects are significant if the ninety-five per cent confidence interval does not include the coefficient 0. Negative coefficients indicate that the number of errors decreases as the value of the predictor increases. By far the largest effect is SNR. The significant interaction between SNR and system indicates that HTS benefits more from high SNRs than F-USEL. The smallest of the three effects, system, indicates that HTS mostly outperforms F-USEL.

Since there are no single point estimates for group-level predictors, but a set of estimates for each group, these predictors are typically summarized using the standard deviation of the estimates—the higher the standard deviation, the stronger the group-level variation. For our participant-level groups of scores, the standard deviation of the intercept is 0.75, for sentence-level groups, it is 0.39, so participants account for more variation in the scores than differences in the sentences we used.

3.5.3 Discussion

Our results confirm the utility of carrying out further studies into the intelligibility of modern TTS systems in noise. We established the baseline set of data for the two systems in fluctuating noise and, as expected, our results showed that intelligibility decreased as the SNR worsened. However, we had also expected one system to outperform the other across the whole SNR range in the same way that Lancaster et al. [2004] found that unit selection outperformed formant synthesis at all SNRs. Instead, our data show a puzzling interaction between SNR and system type. We only observed an advantage of HMM-based HTS over unit-selection-based F-USEL for non-negative SNRs; for negative SNRs, performance was similar. It is not clear why the intelligibility of HTS should deteriorate more than the intelligibility of F-USEL. Possible candidates for an explanation include the richer acoustic detail preserved in unit selection, which might be used by listeners in difficult listening conditions, and the algorithm used to generate the glottal component of HTS, which introduces a slight buzz into the signal. Furthermore, the performance curve shown in Figure 3.1 also changed at 0 dB, becoming markedly steeper. When plotting SNR versus intelligibility, the shape of the typical curve is sigmoidal, with a well-defined floor, a ceiling, and maximal slope at the fifty per cent intelligibility point, the SRT. Since average error rates are still below 2.5 at −10 dB, it appears that we have covered slightly less than half the typical curve, thus indicating SNRs warranting further investigation.

We successfully trialled Matrix sentences, simple declarative sentences with a fixed structure, unlike many intelligibility studies, which rely on SUS material. Despite the relative simplicity of these sentences, we obtained meaningful results. Since the vocabulary is highly controlled, it is possible to administer tests based on the Matrix sentences using a multiple-choice paradigm. Another advantage of the Matrix sentences is their use in audiological testing [Wagener, 2009]. Using the same standardized material for determining SRTs for human and synthetic speech not only makes it easier to compare different types of systems, but also makes it easier to link a person's performance on an intelligibility test to their hearing. This raises a number of interesting questions about the differences between Matrix sentences and SUS, and different paradigms for using Matrix sentences in intelligibility tests. A comparison of Matrix sentences with SUSs was presented in Section 3.2.

Our use of the ICRA noise was vindicated by our results, once again allowing intelligibility testing of synthetic speech to be bridged to audiological testing. However, since the type of noise is known to affect the SRT [Rhebergen and Versfeld, 2005], the generalizability of our results to different kinds of noise typical of the home had not been established. Furthermore, the stimuli for this experiment only contained

synthetic speech, so no comparison with natural speech could be made. We remedied both these situations in future experiments by including natural speech and by using background noises identified as being prevalent amongst participants on AMT. We also aimed to cover even lower SNRs and determine the corresponding performance floor in future work.

Our first step, though, was to establish AMT as a platform for carrying out further experiments.

3.6 Amazon Mechanical Turk experiment

In this experiment, our primary goal was to establish whether the intelligibility of speech synthesis systems can be assessed using crowdsourcing, more specifically the crowdsourcing platform Amazon Mechanical Turk.

In intelligibility tests, we measure to what extent participants can reproduce the content of one or more utterances produced by a given speech synthesis system. The degree to which participants are successful depends on many factors apart from the synthesis system itself, such as the participant's hearing [Wolters et al., 2007], the listening environment [Venkatagiri, 2003], and the participant's familiarity with the languages and voices used. In laboratory settings, we can isolate the effect of system quality by controlling most of these confounding factors. When crowdsourcing, this is more difficult. We have no control over the circumstances under which participants work, and we do not know to what extent they are honest about the information they share.

In speech and language technology, AMT had predominantly been used for annotating and transcribing in order to create rich data sets [Callison-Burch and Dredze, 2010]. As in annotation and transcription, we have a ground truth in intelligibility testing—the words that were spoken by the speech synthesis system. However, there are two important differences. First, the more intelligible a system is, the smaller WERs will be, so distance to the gold standard cannot easily be used to weed out bad participants. Because there are many legitimate sources of inter-participant variability, we need to find a measure that allows us to separate representative participants and outliers. Secondly, we are not only interested in absolute intelligibility scores, but we also want to rank systems in terms of significant differences in intelligibility [King and Karaiskos, 2009]. So, whether AMT is a useful venue for administering intelligibility tests does not just depend on the absolute scores generated, but on whether these scores preserve the relative order of synthesis systems.

We hypothesized that the results of AMT participants would yield similar rankings of the intelligibility of speech synthesis systems compared to the results obtained in

the laboratory with students, but that their absolute WERs would be higher. Callison-Burch and Dredze [2010] give a number of reasons why the data from AMT may be of a lower quality, such as the task being too complex or the instructions not being clear enough. In our case, we expected that background noise and other distractions—which are eliminated in a laboratory environment—would reduce performance.

Our secondary goals were to investigate the effect of participant-specific conditions such as background noise and hearing on intelligibility, and to develop a method for screening out unreliable respondents.

Since our focus was on exploring and validating the methodology, we set up a two-part experiment for which the outcomes can be predicted relatively well from the literature, the first comparing two systems with a US English voice and the second comparing two systems with a UK English voice. In each pair we used the same speaker to minimize speaker quality variation. In both cases, we expected HTS to be significantly more intelligible than the other system.

3.6.1 Method

For the part using the US English voice, we opted for the *KAL*[5] voice built with the diphone system that is part of the standard *Festival* installation and the HTS version that was built at CSTR for demonstration purposes. The HTS version (2007) was built using speaker-adaptive HMM synthesis [Yamagishi et al., 2009]. A total of 523 sentences taken from the Carnegie Mellon University (CMU) *KAL* Communicator database were used to adapt a basic model, which was mixed-gender and had been trained on US English data. Since HTS tends to score very well in comparative tests, we hypothesized that the HTS system would be more intelligible than the diphone.

For the UK English voice part, we chose the *Nick* voice with the systems reported in Section 3.5.1, that is, F-USEL and HTS. Because we had found HTS significantly more intelligible than F-USEL for high SNRs, we hypothesized that HTS would be significantly more intelligible than F-USEL. The *Nick* voice in F-USEL is based on a total of over ten hours of recordings conducted over two years. The HTS version was trained on around 7000 sentences (9.5 hours) of speech taken from the original recordings and, as enough training data is available, uses only speaker-dependent HMMs [Zen et al., 2007].

We synthesized 50 SUSs for each of the four system and voice combinations. These sentences are a subset of the 100 that were used in the *Blizzard Challenge* 2009 [King and Karaiskos, 2009]. They consisted of twelve commands, eight questions, ten statements with a relative clause (complex statements), and twenty statements with

[5] http://www.cstr.ed.ac.uk/projects/festival/onlinedemo.html

no minor clauses (simple statements). The sequence of sentences was randomized once for all four system and voice combinations. One of the commands occurred twice, once in the middle of the sentence list and once towards the end, so that we could test for participants gaming the system and for any learning effect.

Participants were recruited from two sources, the University of Edinburgh student population (lab participants) and AMT (AMT participants). All participants were required to be native speakers of US English to avoid effects of dialect [Lehiste and Peterson, 1959]. Both sets of participants were presented with an identical set of web pages. Those pages requesting demographic information had no defaults set and the layout was designed so that no answer was easier to select than another. Responses to the speech stimuli were typed in by the participant.

Lab participants were recruited through the University of Edinburgh's student employment service and paid £5 for their participation. The experiment was conducted in a quiet meeting room and participants listened to stimuli over Beyerdynamic DT770 PRO 250Ω headphones fed from an Apple MacBook Pro. We recruited a total of 20 participants, 5 for each combination of system and voice. Each participant heard only one combination, thus allowing us to present numerous sentences per combination and to keep the AMT task at a manageable size. Females accounted for 80 % of participants and all were aged between 18 and 29—highlighting just how unrepresentative of the population student participants can be. On the HHIA-S, 16 participants scored 0, 1 scored 4, and 1 female participant scored 28, well above the cut-off for potential sensorineural hearing loss. This participant did not report any problems with her hearing in the free comment field and had the second-best average WER out of the five students who listened to that particular combination of system and voice. No participant wore hearing aids.

Participants recruited through AMT were paid US$1 for the task, with the time for completion set to one hour. We restricted the experiment to US participants and required participants to wear headphones and be native speakers of US English. After initial slow recruitment, the task was re-released every day at a time that roughly corresponded to morning in the US. This led to an average of 20 new completed assignments per day. Out of a total of 229 AMT participants, 73 % completed the entire task, 11 % completed the demographic questionnaire, but failed to transcribe all 50 sentences, and 16 % did not complete any subset of the task. Completion rates were spread evenly across system and voice combinations (Fisher's Exact test, $p = 0.83$), so it does not appear that participants withdrew simply because a particular system was too difficult to understand.

In order to comply with the privacy policy of AMT, we allowed AMT participants to opt out of almost all the demographic questions except for current location within

the US, country of birth, and dialect of English, which we used to filter out people who were not native speakers of US English and who were not born in the US. Only 6 of our participants were born outside the US; of these, 1 was born in New Zealand, and the others were born in places not on our list of English-speaking countries. The New Zealand native also reported being a native speaker of New Zealand English, while the others all stated that they were native speakers of US English. Finally, we excluded two additional participants with a mean WER above 0.9. One of these had trouble playing the stimuli, but the other failed to mention any problems with sound. This left us with a total of 159 participants. The WER criterion appears to be a good method for filtering out AMT participants who did not complete the task conscientiously. The AMT participant with the highest mean WER in the remaining data set had a mean WER of 0.61 and scored 100 per cent WER on only 5 sentences. The AMT participant with the next worst mean WER (0.48) scored 100 per cent WER on 2 out of a total of fifty sentences, only 1 other had two sentences with 100 per cent WER, and 8 others had only one sentence with 100 per cent WER.

Whilst all 159 AMT participants specified their occupation, 1 chose not to report age; 1 did not report a gender; and 4 did not specify their level of education. Of the remainder: 52 % of AMT participants were female; 47 % male; 58 % were aged between 18 and 29, 33 % between 30 and 49, and only 8 % were aged 50 or older. Age groups and genders are distributed evenly across system and voice combinations (age: Fisher's Exact test, $p = 0.432$, gender: Fisher's Exact test, $p = 0.797$). Those 25 % who described themselves as computer scientists are distributed equally across all four combinations ($\chi^2(3) = 1.88$, $p = 0.597$). Synthetic speech was listened to at least weekly by 17 %. Full- and part-time work accounted for 42 % of reported occupations, 25 % were students, 28 % were homemakers, retired, or fell into a category not covered by our alternatives. While 74 % had a college education or a bachelor's degree. 11 % had a postgraduate degree and 12 % had only completed high school. Although 2 had not completed high school, one of these appears to have been a high school student.

We assessed background noise with three questions: level of background noise (quiet all of the time, quiet most of the time, equally quiet and noisy, noisy most of the time, noisy all of the time); type of noise (not applicable, radio/television (TV), chat, music, traffic, domestic); and character of noise (not applicable, constant, fluctuating, short bursts). Of all AMT participants, 54 % reported that their environment was quiet all of the time, for 38 %, it was quiet most of the time. Only 2 % reported a noisy environment most or all of the time. The most frequent type of noise reported by the AMT participants who heard a background noise was radio/television (40 %), followed by traffic (22 %) and chat (16 %). For 38 % the noise came in short, isolated bursts, for 25 % it was constant, and for 26 %, it fluctuated. Noise types, levels, and

sources were evenly distributed across all four combinations of system and voice.

None of the AMT participants had been fitted with a hearing aid. A number of them (19) scored 10 or higher on the HHIA-S, which means that they possibly have a sensorineural hearing loss. In the free comments, an additional 7 who scored low on the HHIA-S mentioned hearing problems, either due to hearing loss or in specific situations. However, none of these participants had the highest mean WER for the stimuli they heard and only 3 of them were in the top 5 for their group.

3.6.2 Results

In the following, we retain WER for descriptive statistics, but for statistical modelling, we converted it to the number of errors that were made on each sentence. Word error rate itself is not normally distributed; 34 % of all scores are 0 and 63 % are below 0.2, and 84 % are below 0.4. By replacing WER with the corresponding number of errors, we obtain an outcome variable that can be approximately characterized using the negative binomial distribution that we used in our baseline experiment.

We modelled the effect of speech synthesis system type on the number of errors made using GLMMs [Gelman and Hill, 2006; Baayen, 2008] with the formula:

$$\text{Error} \sim \text{System} * \text{SentenceType} + (1 + \text{System} \mid \text{Sentence}) + (1 \mid \text{Participant}) \quad (3.4)$$

Individual-level effects were synthesis system (diphone versus HTS for the *KAL* voice; F-USEL versus HTS for the *Nick* voice), and type of sentence (command, question, statement, complex statement). We also included a term for the interaction between system and sentence type. We added two sets of group-level predictors, a sentence-level term and a participant-level term. The participant-level term only consisted of an intercept, which reflects individual differences in performance. We can use these intercepts to identify participants who have particular problems with the material. The sentence-level term consisted of an intercept, which reflects differences in difficulty between sentences, and a slope for speech synthesis system. If the slope for a given sentence is negative, participants are less likely to make errors on that particular sentence when a particular synthesis system is used; if the slope is positive, participants are more likely to make errors.

Models were fitted using the *R* [R Development Core Team, 2009] package *lme4* [Bates and Maechler, 2009]; for Kruskal-Wallis, Wilcoxon, and Spearman tests, we used the package *coin* [Hothorn et al., 2008]. Many p values are very small, even though the actual improvement in model fit is relatively small, so we limited our reporting of them to 0.001, even if the actual figure is smaller.

As we had hypothesized, the mean WER of AMT participants across all four

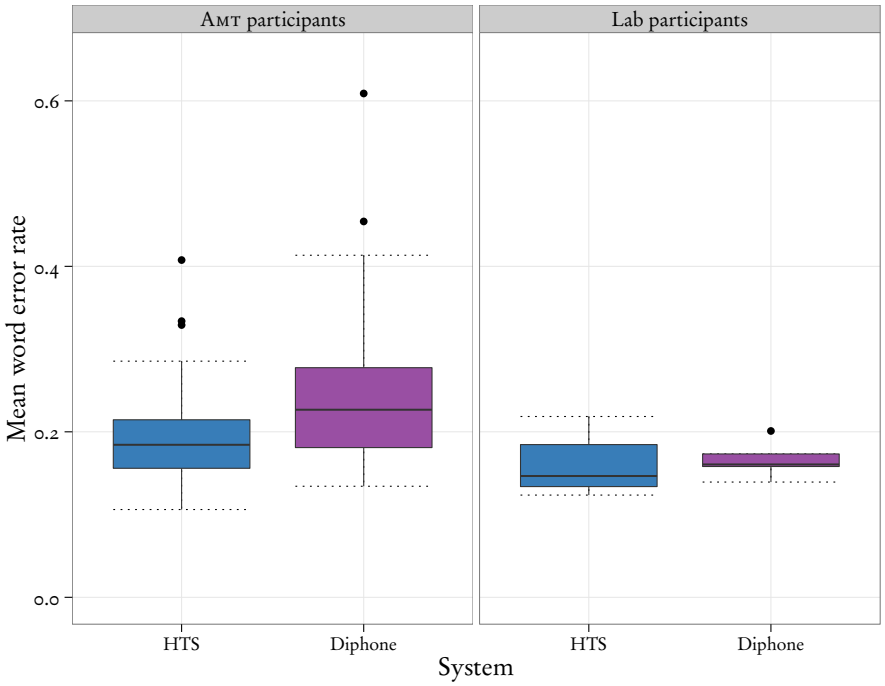

Figure 3.4: *Mean WERs by* KAL *system with* AMT *and lab participants*

combinations was 0.2 compared to 0.13 for the lab participants—some 34 % higher. If we consider only those AMT participants who identified themselves as students, the difference between the two conditions is the same, with a mean WER of 0.19 for student AMT participants. Both differences are statistically significant at the $p < 0.001$ level (Wilcoxon test). When comparing scores on the first and the last ten sentences, the AMT participants show significant learning effects (Wilcoxon test, $Z = 5.7$, $p < 0.001$), but not the lab participants (Wilcoxon test, $Z = 0.67$, $p = 0.501$). None of the groups managed to improve WERs on the command that is repeated.

While absolute WER scores from AMT are much higher, relative differences between systems are either preserved or enhanced. Figure 3.4 summarizes WER for the two systems based on the *KAL* voice, and Figure 3.5 shows WER for the *Nick* voice. Both figures are box-and-whisker plots, with the boxes representing the interquartile range and the whiskers 1.5 ×the interquartile range. Dots are outliers; solid lines indicate medians.

In order to examine the effect of speech synthesis system, type of sentence, and the interaction between sentence type and the combination of system and voice, we removed all variables that contained the predictor to be tested from the fitted model. So, when testing for the effect of system and sentence type, we removed both the variable itself and the interaction term, because otherwise part of the effect of the removed variable would have been captured in that term. We then assessed the difference between full and reduced models using ANOVA and the χ^2 test for

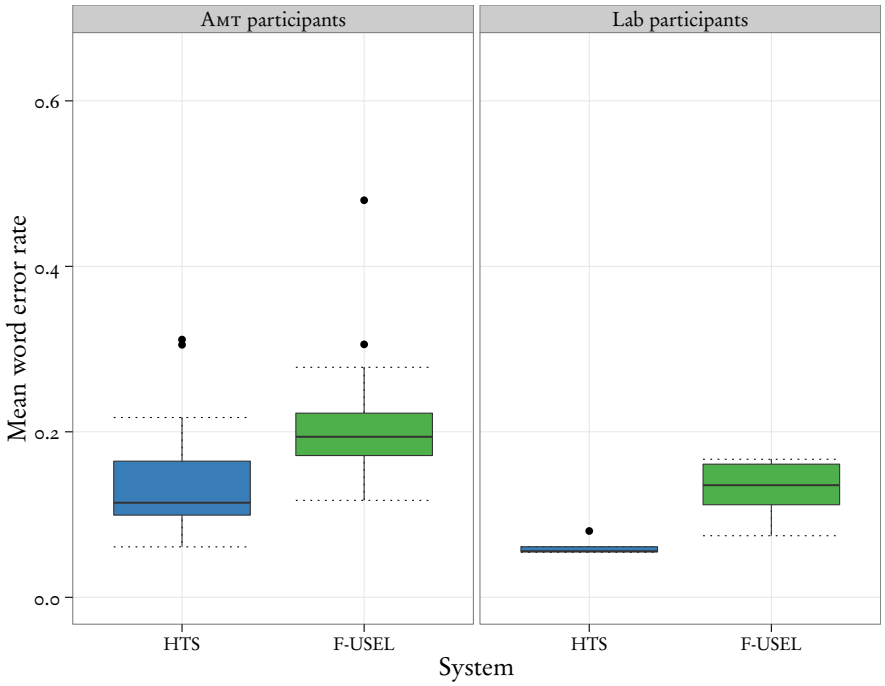

Figure 3.5: *Mean* WERs *by* Nick *system with* AMT *and lab participants*

Table 3.3: *Significance of individual-level predictors (*ANOVA *model comparison, χ^2 test) for the* GLMM *of the* AMT *experiment*

	***Nick* Lab**		***Nick* AMT**		***KAL* Lab**		***KAL* AMT**	
Predictor	**AIC**	***p***	**AIC**	***p***	**AIC**	***p***	**AIC**	***p***
Baseline	918	—	9558	—	1286	—	11052	—
Sentence type	920	0.026	9565	0.004	1283	0.184	11055	0.014
System	935	0.001	9580	0.001	1282	0.362	11660	0.001
System x Sentence type	925	0.004	9569	0.001	1284	0.227	11057	0.010

establishing significance. The results are summarized in Table 3.3.

For each combination of experiment and participant group, we give the AIC of the baseline model with all predictors, followed by the AIC of the models that result when one of these predictors is removed, and the probability that the difference between the original and the reduced model is significant. While the results from the lab participants did not differentiate between the two systems used with the *KAL* voice, clear differences emerge from the data collected from AMT. Looking at the box-and-whisker plots in Figure 3.4, we see that this is due to a few of the laboratory students who had particular problems understanding the HTS version of the *KAL* voice (the upper end of the box-and-whisker plot is much bigger than the lower end).

There are also significant effects of sentence type and interactions between sentence type and system. This is illustrated by Table 3.4, which gives mean WERs by sentence type and combination of system and voice, calculated from AMT participants'

Table 3.4: *Mean WERs for sentence types for all system/voice combinations*

	Nick		*Kal*		
Sentence type	**HTS**	**F-USEL**	**HTS**	**diphone**	**TOTAL**
Command	0.20	0.20	0.25	0.17	0.21
Complex statement	0.13	0.22	0.21	0.25	0.20
Question	0.07	0.17	0.17	0.25	0.17
Statement	0.12	0.21	0.17	0.29	0.20
TOTAL	0.13	0.20	0.20	0.25	0.20

responses. Commands are easiest for people who heard the diphone version of *kal*; for all other sentence types, *Nick* HTS gives the best results.

Figures 3.4 and 3.5 reflect substantial variation in AMT participants' WER scores. In order to find out whether participant and environmental characteristics affected scores, we quantified the effect of exposure to synthetic speech, age group, gender, background noise levels, and background noise type on participants' mean scores using a linear model. (Since the mean scores follow a log-normal distribution, we predicted the logarithm of this variable.) An ANOVA showed no significant effects ($R^2 = 0.16$, $F(17, 141) = 1.63$, $p = 0.064$). Nevertheless, looking at the residual plots, we were able to identify three individuals who performed particularly well and one person who performed much worse than expected—this is the AMT participant with the highest overall mean score.

We can do better than modelling mean scores, however. The statistical model contains a term where separate intercepts are fitted for each participant. These intercepts describe the overall WER trends for each participant after global effects of speech synthesis system and sentence type have been taken into account. The histograms for *kal* and *Nick* (Figure 3.6) show several clear outliers with very large intercepts. These correspond to AMT participants whose individual mean score is more than two interquartile ranges above the mean for their respective group, which brings us back to our original graphs of mean scores—such AMT participants will also be represented as circles in a typical box-and-whisker plot.

3.6.3 Discussion

To the best of our knowledge, this was the first time speech intelligibility was measured using a large scale crowdsourcing, or similar, service to recruit participants. Although absolute WERs are much worse than in the laboratory situation—probably because of background noise and distractions, as we had speculated—the AMT results reflect relative differences in intelligibility fairly well. For many applications,

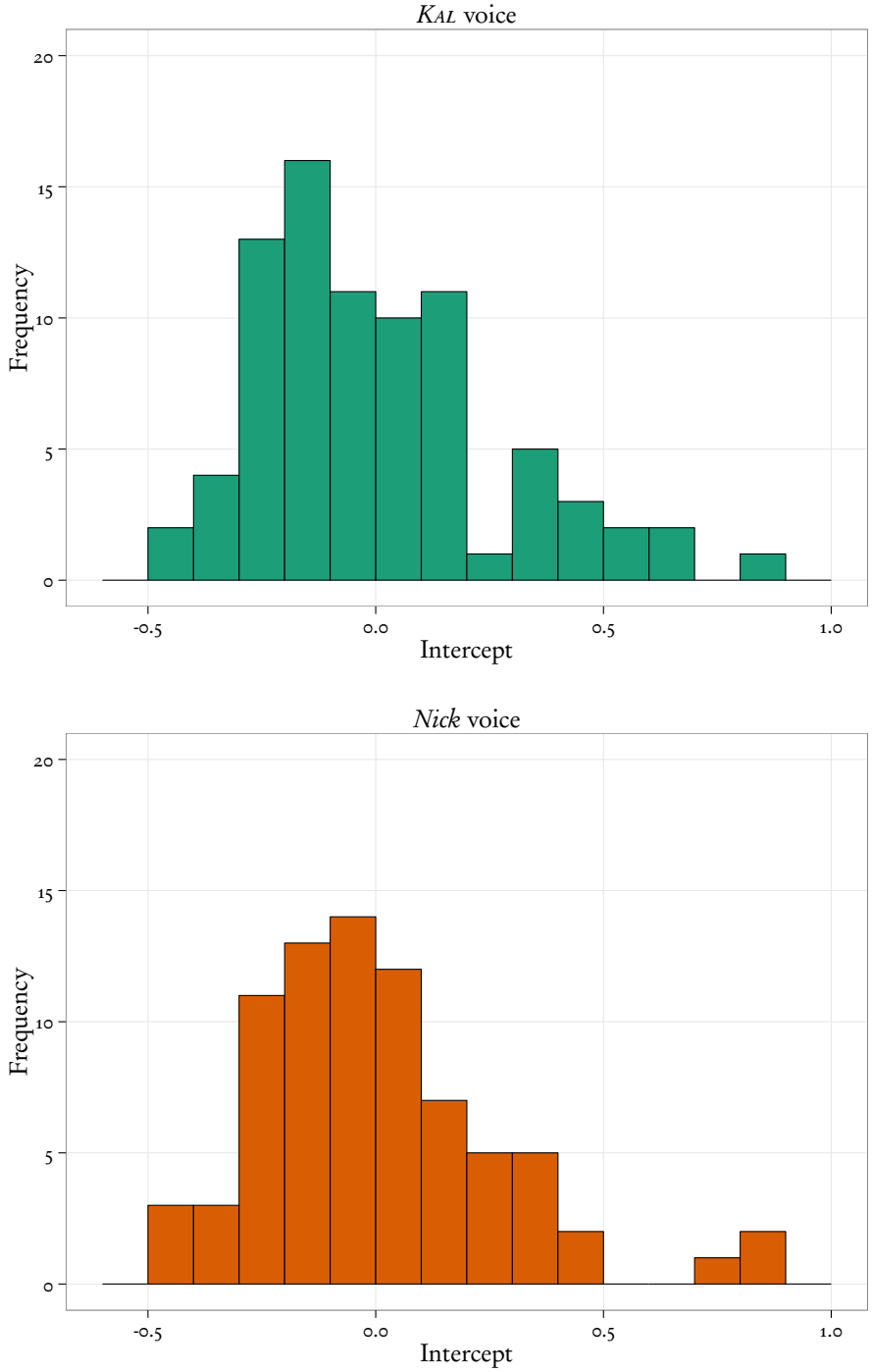

Figure 3.6: *Distribution of participant-level intercepts for* KAL *and* Nick *voices*

for example when a new speech synthesis approach needs to be compared to older technology or to natural speech, this will be sufficient, but not if we want to know whether it is possible to understand synthetic speech perfectly. Participants who perform less well than expected can be identified through simple box-and-whisker plots; these results agree well with the outcomes of GLMMs where separate group-level intercepts are fitted for each participant.

The result of our experiment using the *Nick* voice is as expected, while the outcome of the comparison of the *KAL* systems is somewhat surprising, given that HTS generally yields very high-quality synthetic speech. The poor performance of *KAL* HTS may be due to the lack of material available for adapting the average speaker model. The results on the *KAL* systems illustrate the particular advantage of AMT. For small effect sizes, many participants are needed to obtain sufficient power, especially if the comparison is between subject. Thus, the AMT data were able to reveal a clear difference between the diphone and HTS systems, which had not emerged from the laboratory study due to the small number of participants. We also showed that it is worth including the type of SUS in the analysis, because difficulty varies by type, and some systems seem to perform better on certain types of sentences, such as, statements rather than commands.

All AMT participants who completed the task met the recruitment criteria, except for one who misread the instructions and five native speakers of US English who had not been born in the US. The percentage of sentences that are completely wrongly transcribed (100 % WER) appears to be a good criterion for identifying people who enter random words instead of carefully listening to the sentences. We were somewhat concerned by the high drop-out rate. Around 16 % of all AMT participants failed to complete any part of the task. We did not ask non-completers what their main problem was, but from previous experience with Internet listening experiments, we suspect that playing sound in the participant's browser was the main culprit.

While the sample of AMT participants we recruited is not representative of the population of the US [Ipeirotis, 2010], it is far more diverse than the student and expert samples which are typically recruited for listening experiments. Overall WERs are in line with WERs reported by Novotney and Callison-Burch [2010] for transcription of spontaneous speech. Although the evaluation of AMT in comparing natural speech with synthetic was not conducted in this experiment, it is addressed in Section 3.8. The WERs obtained by the AMT participants in our experiments also vary far more than the WERs of the lab participants. Although we asked all participants to fill in an extensive questionnaire about their listening situation and their hearing, none of these variables was able to cover a sizeable amount of this variation. Indeed, quite a few AMT participants who scored well on the HHIA-S mentioned specific hearing

problems in comments. The analysis of the HHIA-S scores prompted a more detailed study of the use of the HHIA-S for screening in perception experiments [Wolters et al., 2011].

We also suspect that the questions we used to assess environmental noise levels during the test may not have adequately reflected true noise levels and sources. Often, people are exposed to different types of background noise, for example a lorry passing by the window while the radio is on in the background. In our questionnaire, we asked people to highlight only one type of noise and failed to allow for these contingencies. We addressed this shortcoming in subsequent experiments.

We established that AMT is a viable platform for conducting speech intelligibility tests, particularly for investigating methodological issues in intelligibility testing that typically require a large number of participants, such as the choice of sentence material and the effect of learning. Our results supported the use of AMT in future experiments, which we did in our next experiment to evaluate the use of Matrix sentences and, subsequently, in the evaluation of noise and reverberation.

3.7 Matrix sentences experiment

Having established the viability of AMT for hosting listening experiments and, specifically, its ability to match the listening lab in measuring the performance of two systems relative to one another, albeit with more variance, the next step was to ascertain whether equivalent results could be achieved by replacing SUSs with Matrix sentences. To this end, we designed this experiment so that the data collected would match the data collected using AMT for the UK English part (using the *Nick* voice) of the experiment described in Section 3.6, with the exception that the fifty SUSs were replaced by fifty Matrix sentences. Since this experiment was also run on AMT, we continued to collect data that would be useful in assessing AMT and, in particular, data on the types of noises present in participants' listening environments.

3.7.1 Method

The stimuli were generated in the same way as for the baseline experiment described in Section 3.5, that is: the Matrix sentences were chosen randomly using a Latin square; the utterances were synthesized using the two systems used to generate the *Nick* voice files in Section 3.6 (F-USEL and HTS); and the SPL of the files was adjusted to 65 dB.

All participants were recruited through AMT and paid US$1 for their participation. Non-US residents and native speakers of languages other than US English were

excluded from the final analysis. Each participant heard either the F-USEL or the HTS version of the voice. In accordance with the rules of AMT, participants could withhold some demographic information.

A total of 101 AMT participants accepted the task, of whom: 60 % completed all of it; 18 % completed the demographic questionnaire, but failed to transcribe all fifty sentences; and 22 % did not complete any part of the experiment (a 6 % increase over the last experiment). Completion rates were spread evenly across systems (Fisher's Exact test, $p = 0.886$).

Of the 61 who completed all parts of this experiment, 2 of the participants were not born in the US and were excluded, leaving a total of 59 for the remaining analysis. Unlike the previous experiment, no AMT participants were excluded because of their mean WER score, since the AMT participant with the highest WER only scored 0.32 (with no scores at 100 %) and only two 100 % WER scores were recorded across all participants.

Very little demographic information was withheld. Occupation was withheld by only 2 participants. A different 2 withheld their gender, 1 of whom also withheld age. This left 41 % of the participants reporting themselves as working full- or part-time, 25 % as students, and 31 % as homemakers, retired, or in a category not covered by our alternatives. Females made up 54 % and males 42 %; 54 % were aged between 18 and 29, 31 % between 30 and 49, and only 14 % were aged 50 or older. Age groups and genders are distributed evenly across systems (age: Fisher's Exact test, $p = 0.815$, gender: Fisher's Exact test, $p = 0.894$). Listening to synthetic speech occurred at least weekly for 14 % and 25 % described themselves as computer scientists; they are distributed equally across systems ($\chi^2(1) = 0.27$, $p = 0.602$). Having had a college education or a bachelor's degree accounted for 71 %, whilst 10 % had a postgraduate degree and 19 % had only completed high school. No participants reported their environment being noisy most or all of the time and 58 % reported it being quiet all of the time. The remainder reported it as quiet most of the time (34 %) or equally noisy and quiet (8 %).

No participant had been fitted with a hearing aid. Only 4 participants scored 10 or higher on the HHIA-S, which means that they possibly have a sensorineural hearing loss. In the free comments, an additional 3 who scored low on the HHIA-S mentioned minor hearing problems. Only one of these participants had a noticeably high mean WER, which was still only the third highest for that group.

In addition to potential hearing loss (an HHIA-S score of ten or higher) and a reported problem with hearing, we used synthetic speech listening frequency, age group, gender, and whether participants described themselves as a computer scientist as predictors in a linear model. Because the mean WER was not normally

Table 3.5: *Occurrence of background noises among AMT participants of the Matrix sentences experiment. Figures in the grid only include the number of occurrences with another noise, whereas TOTAL additionally includes the noise being reported alone*

	No noise	Construction	Conversation	Music	Radio/TV	Traffic	Other
No noise	—	0	0	0	0	0	0
Construction	0	—	0	0	0	0	0
Conversation	0	0	—	0	2	3	1
Music	0	0	0	—	0	0	0
Radio/TV	0	0	2	0	—	1	0
Traffic	0	0	3	0	1	—	0
Other	0	0	1	0	0	0	—
TOTAL	31	0	8	1	8	11	6

distributed, we used *boxcox* in the *MASS* package [Venables and Ripley, 2002] to calculate a suitable power transformation (0.4), which was then applied to predict mean WER raised to that power. An ANOVA of the resultant model showed no significant effects ($R^2 = 0.17$, $F(11, 47) = 0.89$, $p = 0.556$).

For this experiment, rather than ask for a single main noise, we asked participants to record which of one or more noises they heard. Slightly over half (53 %) of participants reported hearing no noise at all, which—although refuting our intuition that background noise is present more often than not—demonstrates how prevalent background noise is in everyday environments. Table 3.5 summarizes the frequency with which each background noise was reported and its co-occurrence with other noises. Although traffic noise was the single most commonly reported noise, when we consider conversation and radio/TV together, the results tend to support our hypothesis that speech-related noise is the most common. (Although we cannot be sure, it is probable that reports for 'Music'—and possible that 'Other'—also contained speech.) Participants were also asked to rate the overall loudness of the noise using a five-point scale from 1 (no noise) to 5 (very loud). Whilst a linear model showed the effect of noise level on the (transformed) mean WERS was not significant $R^2 = 0.06$, $F(3, 55) = 1.16$, $p = 0.333$, the trend shown in Figure 3.7 is as expected, with an increasing level of noise causing higher WERS.

3.7.2 Results

The first of our concerns with Matrix sentences was that their simple nature could lead to a ceiling effect, particularly, as in this experiment, when used without background

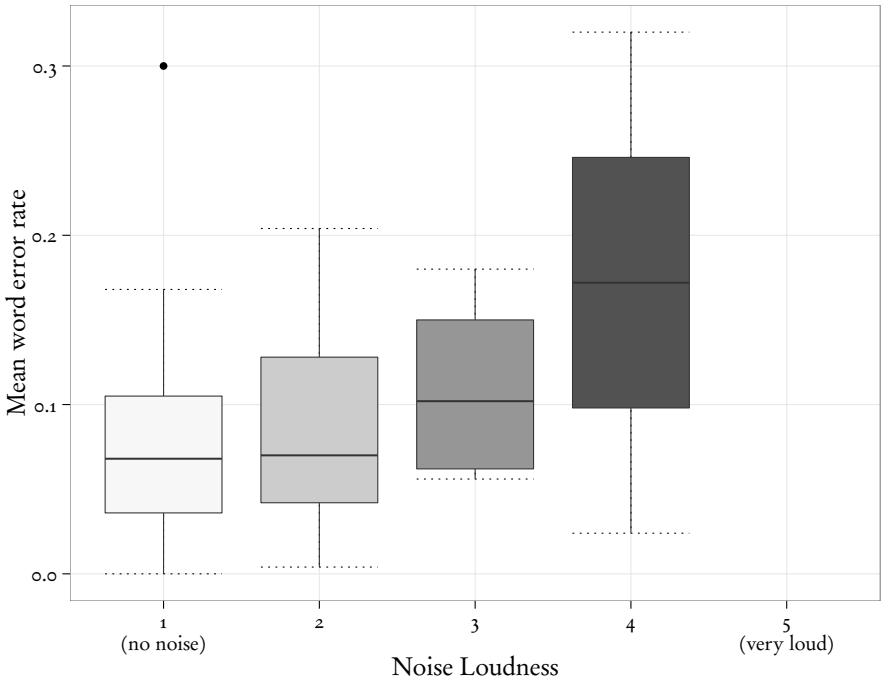

Figure 3.7: *Mean WERs by background noise level perceived by AMT participants in the Matrix sentences experiment*

noise. Inspection of the individual WER scores and the very low overall mean WER (0.09) does suggest a tendency towards a ceiling effect. Indeed, 1 participant achieved a mean WER of zero and no one scored more than a mean WER of 0.32. However, when a histogram of WERs using Matrix sentences is plotted alongside those from using SUSs, from the previous experiment, *using the same bin size* they appear remarkably similar (see Figure 3.8) and the large proportion of very low WERs may just be a symptom of the high intelligibility of modern synthetic speech rather than any failing in the speech materials. However, it should be noted that, because Matrix sentences can only ever have five words, a WER score between 0 to 1 can only take on the values 0, 0.2, 0.4, 0.6, 0.8, and 1; whereas, because SUSs come in lengths of 6, 7, and 8, the resulting WERs can take on nineteen separate values and the actual scores are given in the bottom graph in Figure 3.8 for reference. From this it can be seen that, whilst SUSs appear to be more discriminative, on closer inspection it is clear that the scores to the right of 0 occur in bunches of three, each of which represents the number of errors divided by the three lengths of sentence (6, 7, and 8 words). This in turn means that the WER outcome is dependent on the choice of SUS materials and we have already shown in Table 3.3 that choice of sentence is a significant predictor of a system's performance.

Our intuition was that SUSs introduced more variability into the scores than Matrix sentences, simply because SUSs can include questions, commands, and complex statements and the number of words in each sentence varies. Aggregating the WER

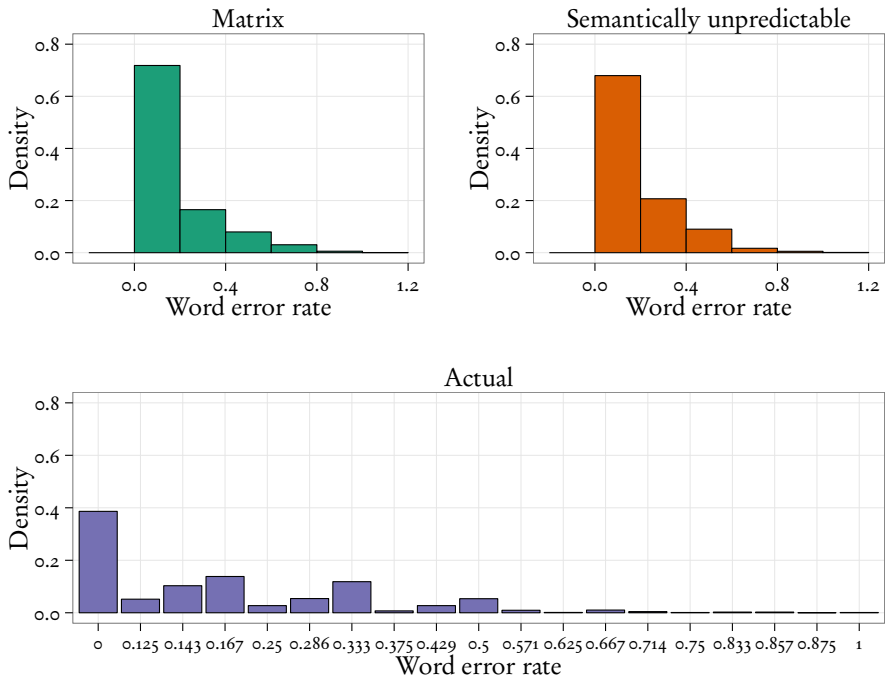

Figure 3.8: *Distribution of WERs with Matrix and semantically unpredictable sentences, using the same bin sizes (top) and actual WERs for SUSS (bottom)*

scores by sentence (rather than by person) reveals a mean of 0.17 and a standard deviation of 0.09 for SUSS whereas for Matrix the figures are 0.09 and 0.07, respectively. While the scores for SUSS are normally distributed, the very low mean for Matrix sentences reflects a non-normal distribution with a high preponderance of scores of 0.

Investigating this further, we built a GLMM for each of the experiments, using the formula in Equation (3.5), to predict the number of errors with the system as an individual-level predictor and the participant and sentence identifiers as group-level predictors.

$$\text{Error} \sim \text{System} + (1 + \text{System} \mid \text{Sentence}) + (1 \mid \text{Participant}) \tag{3.5}$$

We allowed a random intercept for the participant and sentence identifiers and a random slope for sentence. We then plotted the sentence-level intercepts produced by the models. The intercepts reveal the overall WER trend for each sentence after the effects of system have been taken into account, with a higher intercept indicating a higher WER. Figure 3.9 show the intercepts for the Matrix and semantically unpredictable sentences. The intercepts for Matrix sentences do appear to be slightly more tightly bunched around zero, with fewer outliers, but with one very high outlier. However, the graphs are not so different as to enable us to make clear inferences about either set of sentences.

Our second concern with Matrix sentences was the learning effect likely to occur

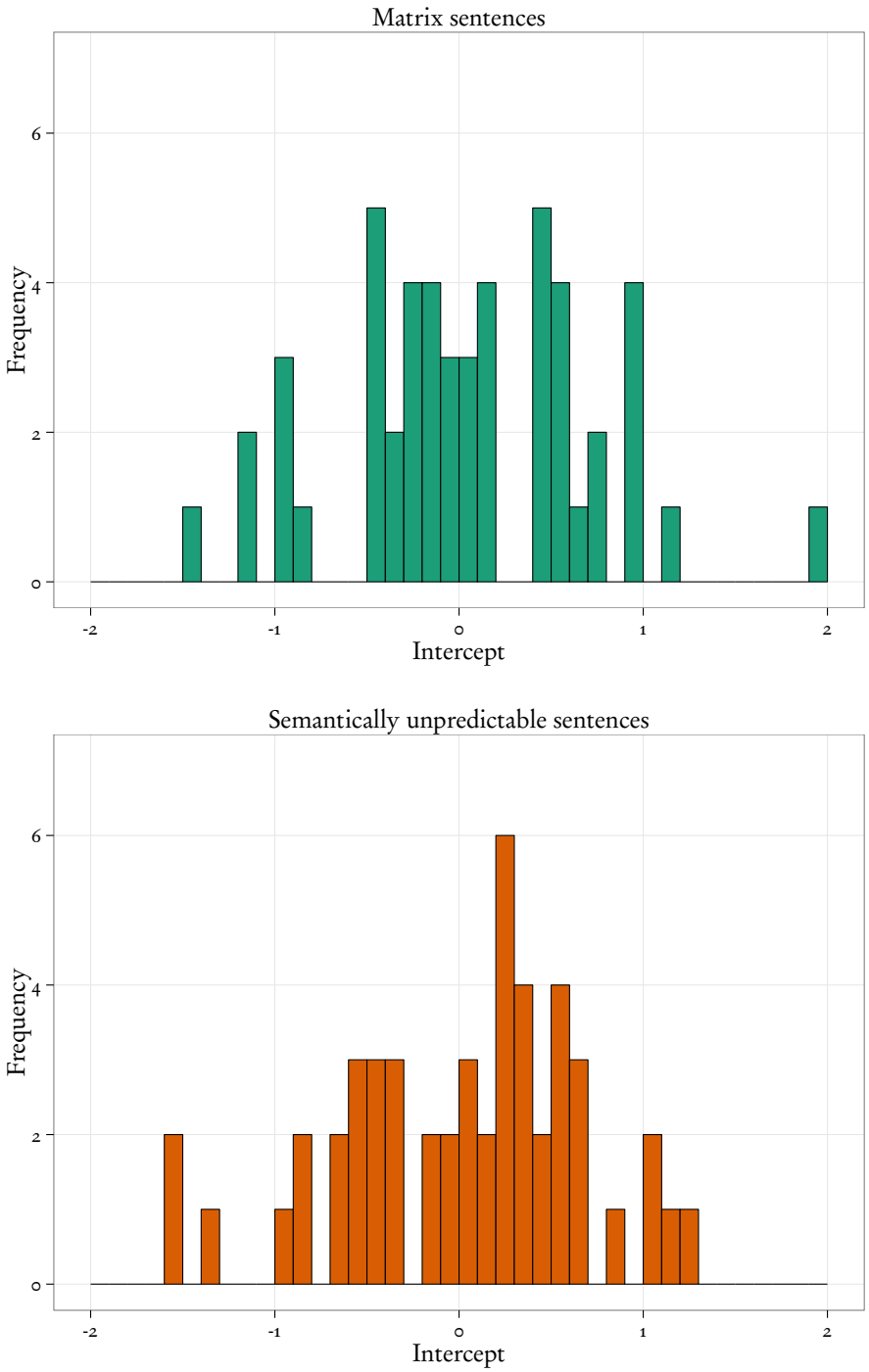

Figure 3.9: *Distribution of sentence-level intercepts for Matrix and semantically unpredictable sentences*

Table 3.6: *Mean WERs for first, penultimate, and last ten sentences (and percentage decrease from first) for semantically unpredictable and Matrix sentences*

	Suss				Matrix			
	HTS		F-USEL		HTS		F-USEL	
Position	WER	%	WER	%	WER	%	WER	%
First	0.165	0	0.199	0	0.067	0	0.190	0
Penultimate	0.145	12	0.198	0	0.054	20	0.134	29
Last	0.148	10	0.201	−1	0.045	34	0.080	58

with the repeated use of a fixed set of words. To assess the impact of any learning effect, we compared WER scores on the first and last ten sentences stratified by synthesis system, as we did with the previous experiment. The result was significant support for a learning effect (Wilcoxon test, $Z = 4.2$, $p < 0.001$).

The normal response to using stimuli that could produce a learning effect is to provide participants with sufficient training that learning ceases. Clearly, this had not happened with the (albeit, limited) training we had provided and we wanted to establish whether learning had peaked at any point in the experiment. To this end, we compared performance on the last ten sentences with the previous (that is, penultimate) ten. Even at this late stage learning was still taking place (Wilcoxon test, $Z = 2.84$, $p = 0.005$).

Although Matrix sentences appear to suffer from a definite and marked learning effect, the same effect was found in the previous experiment when AMT participants listened to SUSS, although not with the lab participants. In order to visualize the extent of the learning effect, we plotted, in Figure 3.10, the WER scores for the first ten sentences, last ten sentences, and the penultimate ten sentences for Matrix and semantically unpredictable sentences. The scale of the learning effect with Matrix sentences is apparent, but not with SUSS and, after further investigation, it seems that the learning effect previously observed with them was a result of the use of the *KAL* voice.

The extent of the learning effect is quantified in Table 3.6, which suggests the effect is more pronounced for the F-USEL system and this is borne out by an ANOVA on a linear model built to predict WER from position (first, penultimate, or last) and the interaction of position and synthesis method ($F(2, 171) = 5.82$, $p = 0.004$).

In spite of the reservations over the potential shortcomings of Matrix sentences, and as Figure 3.11 shows, whilst the absolute levels of WER are lower for both synthesis systems when using Matrix sentences, the relative positions are remarkably similar and both experiments rate HTS as more intelligible than F-USEL. In order to test the significance of the difference between the two systems, using Matrix rather than

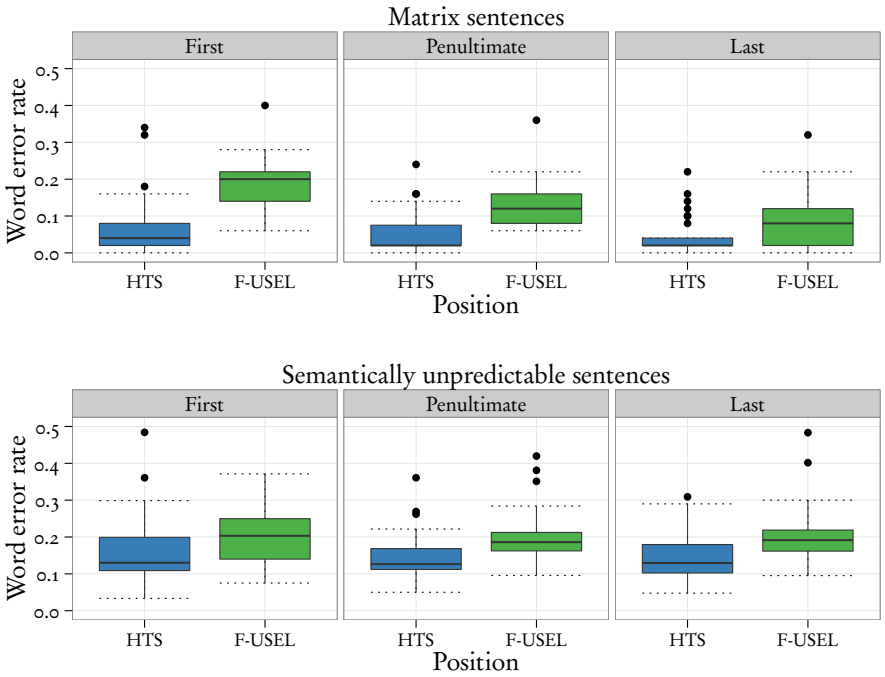

Figure 3.10: *Effect of learning for HTS and F-USEL with Matrix and semantically unpredictable sentences*

Table 3.7: *Significance of individual-level predictors (ANOVA model comparison, χ^2 test) for the GLMMs of the Matrix and semantically unpredictable sentences*

	Matrix		Suss	
Predictor	**AIC**	***p***	**AIC**	***p***
Baseline	4228	—	9565	—
System	4400	0.001	9886	0.001

semantically unpredictable sentences, we used the GLMMs described above. For each experiment, we compared two models: one with the predictor System and one without. Table 3.7 shows the results of ANOVA comparisons of the models from which it can be seen that a significant difference between systems is found with both sets of sentences, even though the absolute WERs for Matrix sentences were somewhat lower.

3.7.3 Discussion

Despite our reservations about the learning and ceiling effects likely when using Matrix sentences without background noise, Figure 3.11 and Table 3.7 demonstrate that results can be obtained with Matrix sentences that match those with SUSS, whilst enjoying the advantages Matrix sentences bring.

The ceiling effect noted with Matrix sentences does not appear materially different

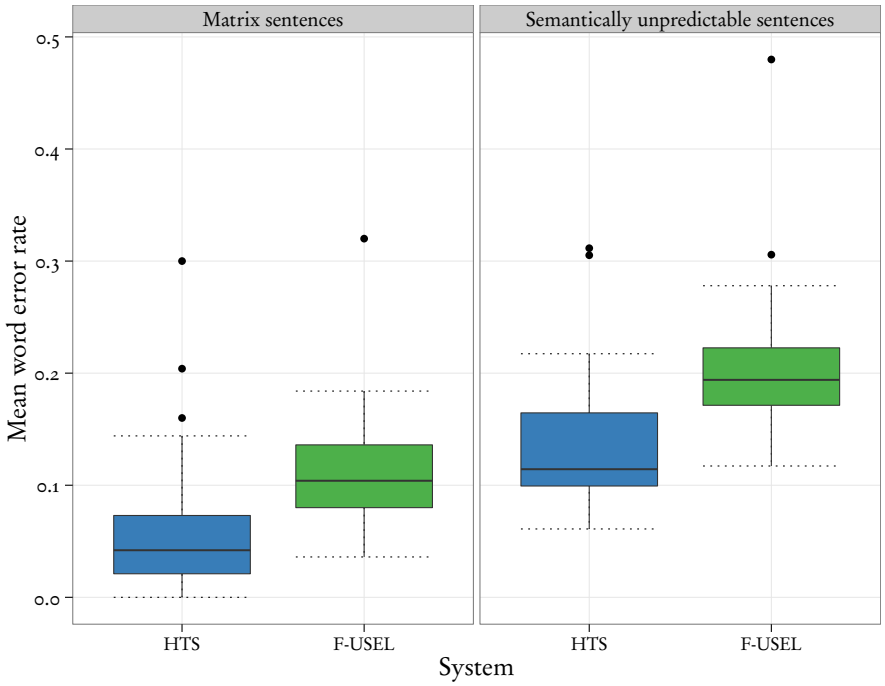

Figure 3.11: *Mean WERs for HTS and F-USEL with Matrix and semantically unpredictable sentences*

from that seen with SUSS in the previous experiment. The wider range of WERS possible with SUSS does not seem to make them any more discriminative in practice and, in any event, the differentials are lost once the WER is converted to numbers of errors. On the other hand, the hypothesized reduction in variability of WER scores resulting from Matrix sentences does not appear to have been realized and they offer no significant advantage over SUSS in this regard.

The learning effect expected with Matrix sentences is significant and is more pronounced than with SUSS. We presume that the reason for this is the predictability inherent in using a relatively closed set of words. Even though the training we provided was somewhat limited, it does not appear, from our results, that further training would make any substantial difference, since learning occurred, at least, up until the penultimate ten sentences. It is not clear why F-USEL should suffer more than HTS with Matrix sentences when it suffered less (in fact, almost not at all) with SUSS.

3.8 New voice and natural speech experiment

The experiments set out in Sections 3.6 and 3.7 had demonstrated the utility of AMT in conducting listening experiments when comparing two synthesis systems and stimulus materials. However, both experiments used the *Nick* voice and neither included natural speech in its comparisons, so it remained to be seen whether our

results were generalizable to other voices and the inclusion of natural speech. In this experiment we sought to test this by using a different voice and comparing stimuli generated synthetically with the same stimuli recorded by the original speaker. The natural speech thus recorded was not used in the training of either synthesis system.

We wanted to compare our results with those obtained from the previous experiments, but were concerned that stimuli created with Matrix sentences in natural speech without noise would have such a ceiling effect as to render the results unusable. Previous experiments had not included natural speech, so its performance on AMT had not yet been compared with that in the lab. Therefore, we reverted to using SUSS and ran the experiment on AMT and in the lab, so that the performance of natural speech on AMT could be established.

3.8.1 Method

The stimuli were generated from sixty SUSS selected from the one hundred used for the *Blizzard Challenge* 2009 [King and Karaiskos, 2009]. The sixty sentences contained fifteen commands, fifteen questions, and thirty statements. In this experiment, all the statements were simple, that is, without a relative clause. The sentence types and three versions of each sentence: HTS, F-USEL, and natural speech (NS) were balanced across three lists. Each participant listened to only one list and, therefore, heard all three systems. Participants were either recruited through the University of Edinburgh's student recruitment service, and paid £5, or through AMT, and paid US$1 for participating. Non-US residents were again excluded as were non-US native speakers of English. The same self-reported information as in Table 3.2 on page 41 was collected from each participant.

A total of 24 lab participants were recruited: 12 male and 12 female. Only two age groups were represented with 96 % of participants being between 18 and 29 and 4 % between 30 and 49. The majority (83 %) were full-time students, with 12 % employed part-time, and 4 % recorded as 'other'. As for education, 8 % reported having finished high school, 46 % some college, 21 % a bachelor's degree, and 25 % a postgraduate degree. Half of the lab participants (50 %) rarely or never listened to synthetic speech whilst 38 % heard it at least twice a year, 4 % heard it at least once a week, and 8 % were not sure. None of the participants wore a hearing aid or scored ten or over on the HHIA-S. The lab part of the experiment was conducted in the CSTR listening booths and with the equipment described in Section 3.5.1.

Of the 110 AMT participants who accepted the task: 55 % completed all of it, 16 % completed the demographic questionnaire, but did not transcribe all sixty sentences, and 28 % did not complete any part of it (another increase of 6 % over the

last experiment). Completion rates were spread evenly across systems (Fisher's Exact test, $p = 0.601$). We excluded participants who were not born in the US, which accounted for 3 of the 61 who completed all parts of this experiment, leaving a total of 58 for further analysis.

As before, no AMT participants were excluded because of their mean WER score, since the AMT participant with the highest mean WER only scored 0.27 (with no individual utterance scores at 100 %) and only 3 WER scores of 100 % were recorded by 3 different individuals.

A total of 4 items of information were withheld: occupation by 3 and gender by 1, which left 38 % of the AMT participants to report themselves as working full- or part-time, 28 % as students, and 29 % as homemakers or in a category not covered by our alternatives. Females accounted for 47 % and males 52 %; 69 % were aged between 18 and 29, 28 % between 30 and 49, and only 3 % were aged 50 or older. Genders (Fisher's Exact test, $p = 0.874$) and age groups were distributed evenly across systems (Fisher's Exact test, $p = 0.048$). Of all the participants, 14 % reported listening to synthetic speech at least weekly and 34 % described themselves as computer scientists; they are distributed equally across all three systems ($\chi^2(2) = 2.04$, $p = 0.36$). Having had a college education or a bachelor's degree accounted for 78 %, whilst 3 % had a postgraduate degree and 17 % had only completed high school. No participant had been fitted with a hearing aid. There were 10 AMT participants who scored 10 or higher on the HHIA-S, which means that they possibly have a sensorineural hearing loss. In the free comments, an additional 2 who scored low on the HHIA-S mentioned minor hearing problems. Only 1 of these participants had a particularly high mean WER score, being the highest for that particular group, but it was not the highest overall and the participant's HHIA-S score was only 12.

No participant reported the listening environment being noisy most or all of the time and 59 % reported it being quiet all of the time. The remainder reported it as quiet most of the time (31 %) or equally noisy and quiet (10 %). Over half (59 %) of participants reported hearing no noise at all, further confirming the results from Section 3.7.1. Table 3.8 summarizes the frequency with which each background noise was reported and its co-occurrence with other noises. In this experiment, conversation rather than traffic noise was the single most commonly reported noise, with 'radio/TV' and 'other' also being relatively common.

Once again, the effect of noise level on mean WERs was not found to be significant from a linear model ($R^2 = 0.06$, $F(4, 53) = 0.88$, $p = 0.482$), when participants were asked to rate the overall loudness of the noise using a five-point scale from 1 (no noise) to 5 (very loud) and the trend shown in Figure 3.12 does not seem as pronounced as that in Figure 3.7 from the experiment that used Matrix sentences.

Table 3.8: *Occurrence of background noises among AMT participants of the new voice and natural speech experiment. Figures in the grid only include the number of occurrences with another noise whereas TOTAL additionally includes the noise being reported alone*

	No noise	Construction	Conversation	Music	Radio/TV	Traffic	Other
No noise	—	0	0	0	0	0	1
Construction	0	—	1	0	0	0	0
Conversation	0	1	—	0	1	0	0
Music	0	0	0	—	0	0	0
Radio/TV	0	0	1	0	—	0	1
Traffic	0	0	0	0	0	—	1
Other	1	0	0	0	1	1	—
TOTAL	34	1	9	1	5	6	7

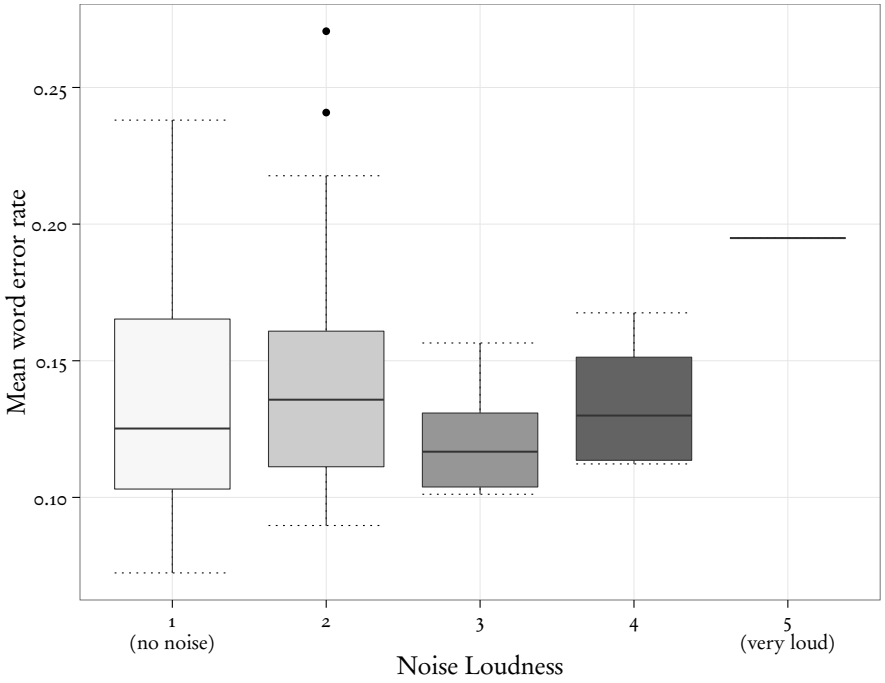

Figure 3.12: *Mean WERs by background noise level perceived by AMT participants in the new voice and natural speech experiment*

Table 3.9: *Significance of individual-level predictors (ANOVA model comparison, χ^2 test) for the GLMM of the new voice and natural speech experiment*

	Lab		AMT	
Predictor	**AIC**	***p***	**AIC**	***p***
Baseline	2383	—	7389	—
Sentence type	2372	0.996	7394	0.007
System	2877	0.001	7481	0.001
System x Sentence type	2376	0.978	7397	0.003

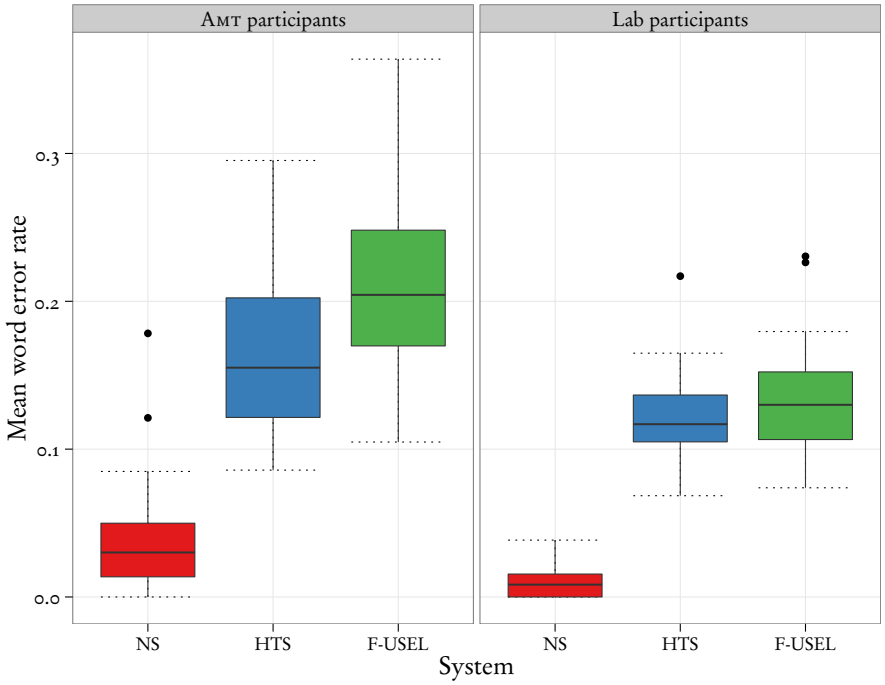

Figure 3.13: *Mean WERs by system with AMT and lab participants*

3.8.2 Results

A linear model was built to predict an appropriate power (-0.1) of the mean WER from potential hearing loss, synthetic speech listening frequency, age group, gender, and being a computer scientist. The linear model showed no significant effects ($R^2 = 0.26$, $F(10, 47) = 1.64$, $p = 0.125$).

As we did in Section 3.6.1 for the Amazon Mechanical Turk experiment, we plotted the mean WERs for each of the systems for AMT participants and lab participants, the results of which can be seen in Figure 3.13. Comparing the performance of HTS and F-USEL using the *Roger* voice in this experiment (Figure 3.13) with that of the *Nick* voice in Section 3.6.1 (Figure 3.5 on page 55), the similarity is clear.

The results for NS for AMT and lab participants are also shown in Figure 3.13. As would be expected, WERs for NS are lower than for both the synthetic systems

Table 3.10: *Mean WERs for sentence types for all system/voice combinations*

	AMT			Lab			
Sentence type	**NS**	**HTS**	**F-USEL**	**NS**	**HTS**	**F-USEL**	**TOTAL**
Command	0.033	0.123	0.280	0.011	0.117	0.171	0.132
Question	0.025	0.163	0.153	0.007	0.105	0.113	0.102
Statement	0.042	0.187	0.211	0.012	0.130	0.129	0.130
TOTAL	0.035	0.165	0.214	0.011	0.121	0.136	0.124

for both sets of participants. As we would also expect, whilst the absolute WER for each system is lower for lab participants, the relative placing of the systems is very similar, although it does appear, from the graph, that there is no significant difference between the WERs for HTS and F-USEL for the lab participants.

In order to investigate this in more depth, we used GLMMs to model errors against system and sentence type (and their interaction) as individual-level predictors and the identifiers for participant and sentence as group-level predictors. The formula used was:

$$\text{Error} \sim \text{System} * \text{SentenceType} + (1 + \text{System} \mid \text{Sentence}) + (1 \mid \text{Participant}) \quad (3.6)$$

We allowed a random intercept for the participant and sentence identifiers and a random slope for the sentence. The results of ANOVA comparisons of models with and without the individual-level predictors and their interaction appear in Table 3.9. As the box plots in Figure 3.13 suggested, both the AMT and lab results show a significant effect of synthesis system, but only AMT shows a significant effect of sentence type or the interaction between sentence type and system. The reason for this is unclear. Perhaps the most likely reason would be that the lab participants are a rather homogeneous group of university students, but creating a model of the AMT results using only those AMT participants who declared themselves as full-time students produces results that closely match those for AMT participants in general. Another candidate is the fact that, whilst lab participants wore studio-quality headphones and listened in a sound-treated booth, the AMT participants wore domestic headphones and listened in their own homes, many of which we know were noisy from participants' own reports and these more difficult conditions could have been enough to cause differentiation between sentence types. Of course, it could just be that having a larger number of participants allowed us to find a small, but significant effect. Unfortunately, as yet, there is no universal agreement on how effect sizes should be calculated from GLMMs.

Table 3.10 and Figure 3.14 show in more detail how mean WERs are spread across system and sentence type for each of the cohorts. Clearly, the pattern of results in the

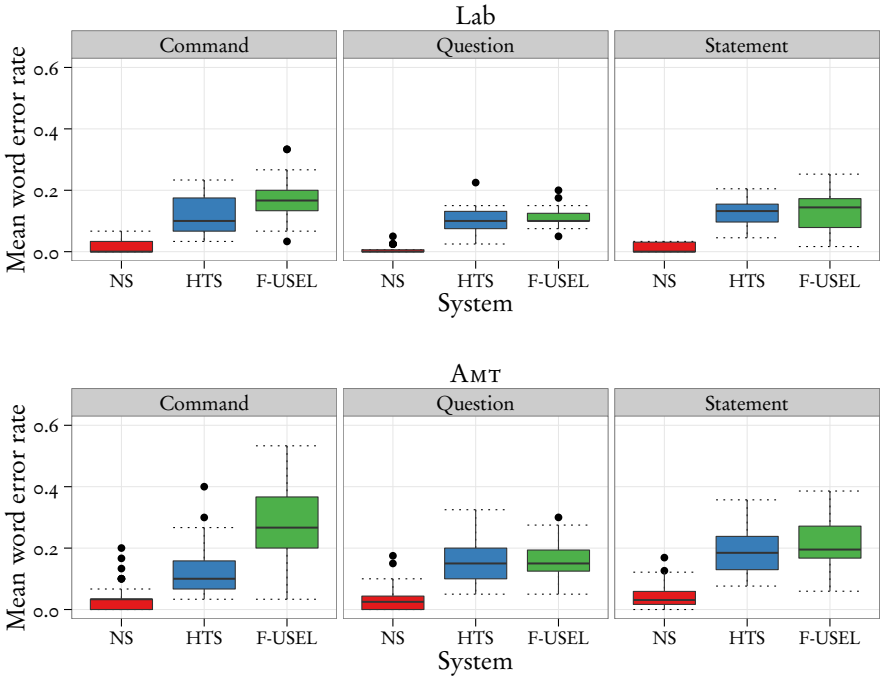

Figure 3.14: *Mean WERs by sentence type and system with AMT and lab participants*

lab is mirrored on AMT. However, there appear to be some anomalies. For example, as might be expected, all sentence types in natural speech have a WER that is very close to 30 % lower for lab participants than for AMT and the position is similar for F-USEL with WERs being 26 % to 39 % lower. The range for HTS, though, is much broader with commands, statements, and questions scoring 5, 30, and 36 % lower respectively. Even more intriguing is that questions score lower for F-USEL than HTS, and statements higher for F-USEL than HTS, with AMT participants, but the opposite is true in both cases with lab participants. (Although the margins are quite small in both cases.) Of most pragmatic use, perhaps, is the fact that, within each experiment, there is less variation between sentence types for HTS than F-USEL.

3.8.3 Discussion

We set out to establish whether our results for the use of AMT were generalizable to another synthetic voice and to natural speech. The results from this experiment clearly confirm our earlier findings and demonstrate that AMT is a suitable alternative to the lab for establishing the rankings of synthesis systems amongst themselves and against their natural speech equivalent.

As in the first experiment using AMT reported in Section 3.6, we found a small but significant effect of sentence type by using AMT that would not have been found by modelling the results from the lab alone. The persistence of the effect across two different voices supports our assertion that sentence type should be an important

consideration when choosing a synthesis system where practicable. In the case of reminders, this could be achieved by framing the reminders using only one sentence type and choosing a synthesis system with a particular bias for that type. For example, a reminder to lock the door could be framed as a command, 'Lock the door.'; a statement, 'The door needs to be locked.'; or a question, 'Have you locked the door?'

The performance of natural speech using semantically unpredictable sentences was very nearly at ceiling for AMT (mean WER, 0.04) and lab (mean WER, 0.01) participants and, because they are inherently more challenging than Matrix sentences, our decision not to use Matrix sentences for real speech in the absence of noise appears to have been vindicated. It would seem that Matrix sentences could not be recommended for listening experiments involving natural speech without noise.

3.9 All Amazon Mechanical Turk experiments

The experiments presented in Sections 3.6, 3.7, and 3.8 all included components conducted on AMT, for which we collected demographic and environmental data. These data give us an opportunity to provide an analysis of participants who opted to take part in the listening experiments.

In all three experiments, many participants (27 %, 40 %, and 45 %, respectively) did not complete the whole experiment, with the number not completing *any* part increasing by about 6 % each time. In order to prevent participants completing any experiment twice, and from taking part in more than one, we recorded their unique identifier and used this to prevent them from taking part again. Unfortunately, it was not possible to check a participant's identifier until the assignment had been accepted, at which point we could prevent access to the experiment. Although the experiment instructions clearly stated that previous participants were excluded, it is likely that some accepted subsequent assignments only to be refused access by us. This would then account for some of those who completed no part of the experiment. In an experiment where participants only hear one system or voice, a situation could arise where numerous participants abandon assignments containing particularly unintelligible speech, thus skewing the results. Our analysis of completion rates for each experiment showed no differences between the groups participants were assigned to, suggesting that no one abandoned an assignment simply because the speech was too difficult to understand. Table 3.11 summarizes all the data collected from those who completed the experiments, excluding the few who did not actually meet the criteria. The data from the lab components of the two experiments that included them are also presented.

We had taken care to minimize the potential for AMT participants to 'game' the

system. For example, we never set default values so that participants could simply click a 'proceed' button without actually make selections themselves and we ensured that no option was more difficult than any other to select. Of course, this did not prevent a participant from selecting the 'wrong' gender or age group, but at least there was no incentive to do so. A side effect of the care we took was that lab participants also could not select options that would enable a more rapid completion of the experiment. For this reason, the results, although self-reported, should be as accurate as possible.

It is clear from Table 3.11 that AMT participants form a more heterogeneous group than lab participants, particularly with regards to age group, education, and occupation. Nearly every lab participant was aged 18 to 29 and a full-time student. Of course this does make for some consistency when assessing synthesis systems over a series of experiments and we wanted to know how much variability there was between cohorts of AMT participants. Each of the factors in the leftmost column of Table 3.11 forms part of a contingency table with the experiments listed along the top of the table, so to assess the variability, we carried out a series of Fisher's Exact tests on these contingency tables. None of the tests found a significant difference between the spread of participants across the factors for the three experiments when they were carried out on AMT. So it seems that AMT cohorts are likely to be similar in make up, certainly in the numbers that we recruited.

The fact remains that listening booths are very effective at eliminating extraneous noise and focussing participants on the task in hand, so it is likely that performance on AMT will always be poorer in absolute terms.

3.10 Conclusion

At the start of the PhD project, we had discovered the paucity of recent research into synthetic speech in noise and reverberation and had identified three recent developments (AMT, Matrix sentences, and the ICRA noise) that could be used in an experimental methodology to enhance the assessment of the intelligibility of synthetic speech in noise and reverberation.

We carried out a series of four experiments with which we sought to establish: baselines for synthetic speech performance and future experimental conditions; whether the lab could be replaced by AMT; whether Matrix sentences could replace SUSS; what the consequences of each of these changes would be; and whether the results were generalizable to another synthetic voice and natural speech.

The ICRA noise was deployed successfully in the baseline experiment to establish the need for research into current synthesis systems by confirming a difference in

Table 3.11: *Demographic data collected from AMT and lab participants from all experiments with an AMT component*

		AMT participants								Lab participants					
		Experiment								Experiment					
		AMT		Matrix		New voice		TOTAL		AMT		New voice		TOTAL	
		n	%	n	%	n	%	n	%	n	%	n	%	n	%
Gender	female	83	52	32	54	27	47	142	51	16	80	12	50	28	64
	male	75	47	25	42	30	52	130	47	4	20	12	50	16	36
	withheld	1	1	2	3	1	2	4	1	0	0	0	0	0	0
Age group	18–29	93	58	32	54	40	69	165	60	20	100	23	96	43	98
	30–49	52	33	18	31	16	28	86	31	0	0	1	4	1	2
	50+	13	8	8	14	2	3	23	8	0	0	0	0	0	0
	withheld	1	1	1	2	0	0	2	1	0	0	0	0	0	0
Education	before school	2	1	0	0	1	2	3	1	0	0	0	0	0	0
	high school	19	12	11	19	10	17	40	14	0	0	2	8	2	5
	some college	72	45	31	53	25	43	128	46	5	25	11	46	16	36
	bachelors	45	28	11	19	20	34	76	28	11	55	5	21	16	36
	masters	15	9	5	8	2	3	22	8	4	20	6	25	10	23
	doctorate	2	1	1	2	0	0	3	1	0	0	0	0	0	0
	withheld	4	3	0	0	0	0	4	1	0	0	0	0	0	0
Occupation	employed full-time	43	27	15	25	17	29	75	27	0	0	0	0	0	0
	employed part-time	24	15	9	15	5	9	38	14	0	0	3	12	3	7
	homemaker	18	11	8	14	6	10	32	12	0	0	0	0	0	0
	other	24	15	9	15	11	19	44	16	0	0	1	4	1	2
	retired	2	1	1	2	0	0	3	1	0	0	0	0	0	0
	student full-time	40	25	15	25	16	28	71	26	20	100	20	83	40	91
	withheld	8	5	2	3	3	5	13	5	0	0	0	0	0	0
Computer scientist	no	119	75	44	75	38	66	201	73	16	80	18	75	34	77
	yes	40	25	15	25	20	34	75	27	4	20	6	25	10	23
Work in speech technology	no	158	99	59	100	58	100	275	100	19	95	23	96	42	95
	yes	1	1	0	0	0	0	1	0	1	5	1	4	2	5
Listening Frequency	at least once a wk	27	17	8	14	8	14	43	16	0	0	1	4	1	2
	at least twice a yr	60	38	17	29	26	45	103	37	5	25	9	38	14	32
	not sure	11	7	8	14	2	3	21	8	3	15	2	8	5	11
	rarely or never	61	38	26	44	22	38	109	39	12	60	12	50	24	55
Headphones	earbuds	66	42	29	49	22	38	117	42	0	0	3	12	3	7
	full ear	30	19	13	22	13	22	56	20	19	95	21	88	40	91
	in ear	19	12	8	14	10	17	37	13	1	5	0	0	1	2
	on ear	44	28	9	15	13	22	66	24	0	0	0	0	0	0
Headphone features	noise cancelling	16	10	3	5	7	12	26	9	0	0	0	0	0	0
	none known	139	87	52	88	49	84	240	87	20	100	24	100	44	100
	sound isolating	4	3	4	7	2	3	10	4	0	0	0	0	0	0
HHIA-S total	10 or above	19	12	4	7	10	17	33	12	1	5	0	0	1	2
	below 10	140	88	55	93	48	83	243	88	19	95	24	100	43	98
Noisiness	quiet all time	86	54	34	58	34	59	154	56	16	80	22	92	38	86
	quiet most time	60	38	20	34	18	31	98	36	4	20	1	4	5	11
	equal noise quiet	10	6	5	8	6	10	21	8	0	0	1	4	1	2
	noisy most time	2	1	0	0	0	0	2	1	0	0	0	0	0	0
	noisy all time	1	1	0	0	0	0	1	0	0	0	0	0	0	0
Browser	Chrome	18	11	11	19	17	29	46	17	0	0	0	0	0	0
	Firefox	92	58	36	61	32	55	160	58	13	65	0	0	13	30
	IE	21	13	5	8	3	5	29	11	0	0	0	0	0	0
	Mozilla	8	5	3	5	1	2	12	4	7	35	0	0	7	16
	Opera	3	2	0	0	0	0	3	1	0	0	0	0	0	0
	other	1	1	0	0	0	0	1	0	0	0	0	0	0	0
	Safari	16	10	4	7	5	9	25	9	0	0	24	100	24	55
Location	East North Central	33	21	4	7	10	17	47	17	0	0	1	4	1	2
	East South Central	9	6	7	12	4	7	20	7	0	0	0	0	0	0
	Mid Atlantic	17	11	11	19	6	10	34	12	0	0	2	8	2	5
	Mountain	9	6	5	8	6	10	20	7	0	0	1	4	1	2
	New England	10	6	3	5	0	0	13	5	0	0	1	4	1	2
	Pacific	32	20	10	17	12	21	54	20	0	0	3	12	3	7
	South Atlantic	23	14	13	22	19	33	55	20	0	0	0	0	0	0
	West North Central	10	6	4	7	0	0	14	5	0	0	0	0	0	0
	West South Central	16	10	2	3	1	2	19	7	0	0	0	0	0	0
	CSTR Lab	0	0	0	0	0	0	0	0	20	100	16	67	36	82
Experience of stimuli	usually all words	5	3	26	44	9	16	40	14	8	40	12	50	20	45
	usually most words	121	76	32	54	45	78	198	72	10	50	12	50	22	50
	very hard	33	21	1	2	4	7	38	14	2	10	0	0	2	5
	TOTAL PARTICIPANTS	159	100	59	100	58	100	276	100	20	100	24	100	44	100

performance between HMM- and unit-selection-based systems and the discovery of an interesting crossover at low SNRs. The range of SNRs it was presented at established a baseline of performance and gave a good indication of those we should use in future experiments.

Crowdsourcing, in the form of Amazon Mechanical Turk, was shown to be a valid alternative to the laboratory in ranking the performance of synthetic and natural speech across two different voices. The wider participant pool provided by AMT makes results more generalizable. Moreover, it was found to have an advantage over laboratory experiments in having the power to find small but significant effects that would otherwise be missed.

We successfully piloted Matrix sentences and found them to be equivalent to SUSs in ranking synthesis systems, albeit at a lower absolute WER.

Our inclusion of questions for AMT participants about environmental conditions allowed us to select background noises for the experiments presented in the next and following chapters.

CHAPTER 4

SYNTHETIC SPEECH IN NOISE

COMMUNICATION using human speech involves at least two parties: a speaker and a listener. The speaker effectively encodes the message to be communicated and the listener decodes it. The encoded message is transmitted through air in the form of sound waves and, just like any other sound (or *signal*), it is subject to constructive and destructive interference by unwanted sound waves (or *noise*). How much the signal is distorted by the noise depends on a number of factors, including: the frequencies and amplitude of each; the hearing ability of the listener; and the listening environment.

4.1 The problem of noise

When listening to any sound, the closer the frequencies in the competing noise are to the signal and the more power (loudness) it has, the harder the signal will be to perceive. The log ratio of the power of the signal (P_{signal}) to the power of the background noise (P_{noise}) is known as the signal-to-noise ratio (SNR), measured in decibels (dB) using Equation (4.1). The SNR may be negative when the level of the background noise exceeds that of the signal.

$$\text{SNR (dB)} = 10 \log_{10} \left(\frac{P_{\text{signal}}}{P_{\text{noise}}} \right) \tag{4.1}$$

In the case of speech, the situation is somewhat more complicated for a number of reasons. Firstly, everyday speech can encompass a range of frequencies from about 50 Hz to 20 000 Hz, so different noises with different spectra will interfere with

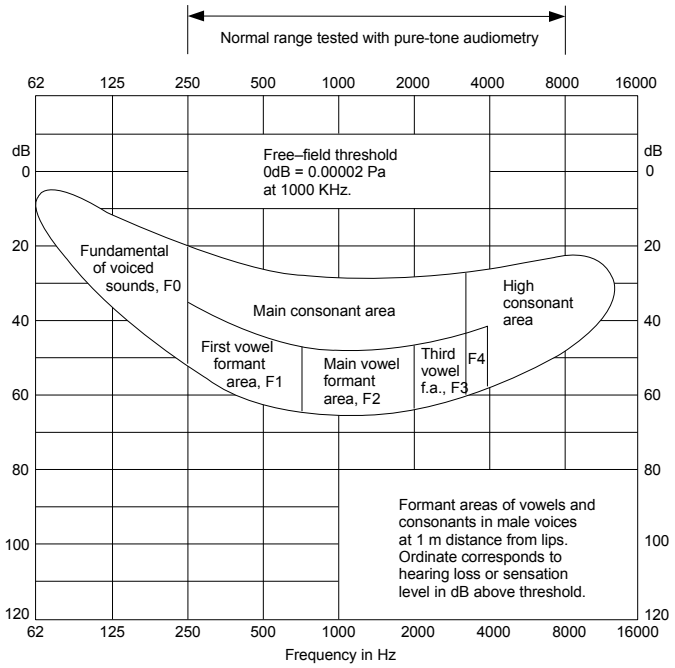

Figure 4.1: *Speech spectrum data schematized in terms of formant areas. The ordinate is sensation level versus free field threshold at 1 metre distance (that is, dB SPL). Reproduced, with permission, from Fant [2005]*

(or *mask*) different parts of the speech [French and Steinberg, 1947]. Secondly, the different frequencies are perceived as having different intensity levels by the human auditory system, even when presented at the same intensity. Thirdly, the different frequencies carry different information, with varying importance to the intelligibility of the message being conveyed.

Figure 4.1 shows that the actual average range of frequencies is somewhat less than the possible extremes and that the core range for speech, that is, the first formant frequency (F1) and the second formant frequency (F2), is limited to about 250 Hz to 2000 Hz. The 'speech banana', as it is sometimes called, also shows a clear separation between vowels and consonants and, indeed, between different types of consonant. Originally, it was thought that consonants contributed more to the intelligibility of speech and the early syllable-level tests were scored on the correct identification of them. More recent work [Kewley-Port et al., 2007] has cast doubt on this assumption, particularly for the intelligibility of whole, meaningful sentences.

Figure 4.1 is laid out as an audiogram and so can be used to estimate what sounds would be heard by a person given their audiogram marked up with the frequencies at which hearing loss has occurred—a fact that forms the basis of the Articulation Index (AI) and its derivatives that were discussed in Section 2.3. More importantly, for our purposes, it can also be used to determine what would *not* be heard in the presence of any noise that has similar frequencies and amplitude. Clearly, the more

similar the noise is to speech, the more it might be expected to interfere, such that another speaker might be very detrimental, a dog whistle not at all, and music might fall somewhere in between.

When the noise is fluctuating, the position becomes even more complicated. It is self-evident that speech has periods of intensity interspersed with pauses or gaps. If one thinks of these as a series of peaks and troughs and fluctuating noise as having a similar structure, it is clear that it is unlikely that, at any given time, the two will coincide. What may not be so evident is that, even when the signal and noise do co-occur, the spectrogram of the additive mixture is almost exactly the maximum of the individual spectrograms (for an explanation of why this is the case, see Roweis [2003]).

Furthermore, human speech naturally has a high level of redundancy: each speech sound is spread over a range of frequencies and persists over time. Together, this means that at the times when speech is dominant (even if these are few and far between) humans are able to catch 'glimpses' of the speech and use their knowledge of the language to 'fill in the blanks' [Cooke, 2003].

Figure 4.2 demonstrates how fluctuating noise interferes with speech. The series of six spectrograms shows an utterance used in the baseline experiment, 'Steven has ten dark desks.', followed by the same utterance with an equal-length section of the International Collegium of Rehabilitative Audiology (ICRA) noise added at SNRs of 10, 5, 0, −5, and −10 dB. The spectrograms are three-dimensional in that they show time on the x axis, frequency on the y axis, and relative amplitude by intensity of colour from blue to red, through cyan and yellow. The ICRA noise is fluctuating in nature and it is particularly noticeable from the second and subsequent images where it has been added to the beginning and end of the utterance where there was no speech. Closer inspection also shows how the speech around the 1 s mark becomes relatively less intense as the SNR becomes more challenging.

Listening to speech in background noise is especially difficult for people with hearing loss and is one of the most common complaints that they express [Plomp and Mimpen, 1979; McArdle and Wilson, 2009]. Indeed, it is recognized by audiologists as one of the first signs of auditory ageing. Much research into human speech in noise has, therefore, been conducted by audiologists focused on measuring how much of the speech an individual can perceive, both as a measure of their impairment and as a measure of the efficacy of new hearing aids and the algorithms that underpin the signal processing they carry out. Most of this work has been conducted in isolation from the speech synthesis research community and we sought to bridge this gap with the use of Matrix sentences.

When communicating in a background noise, it is common for speakers to make

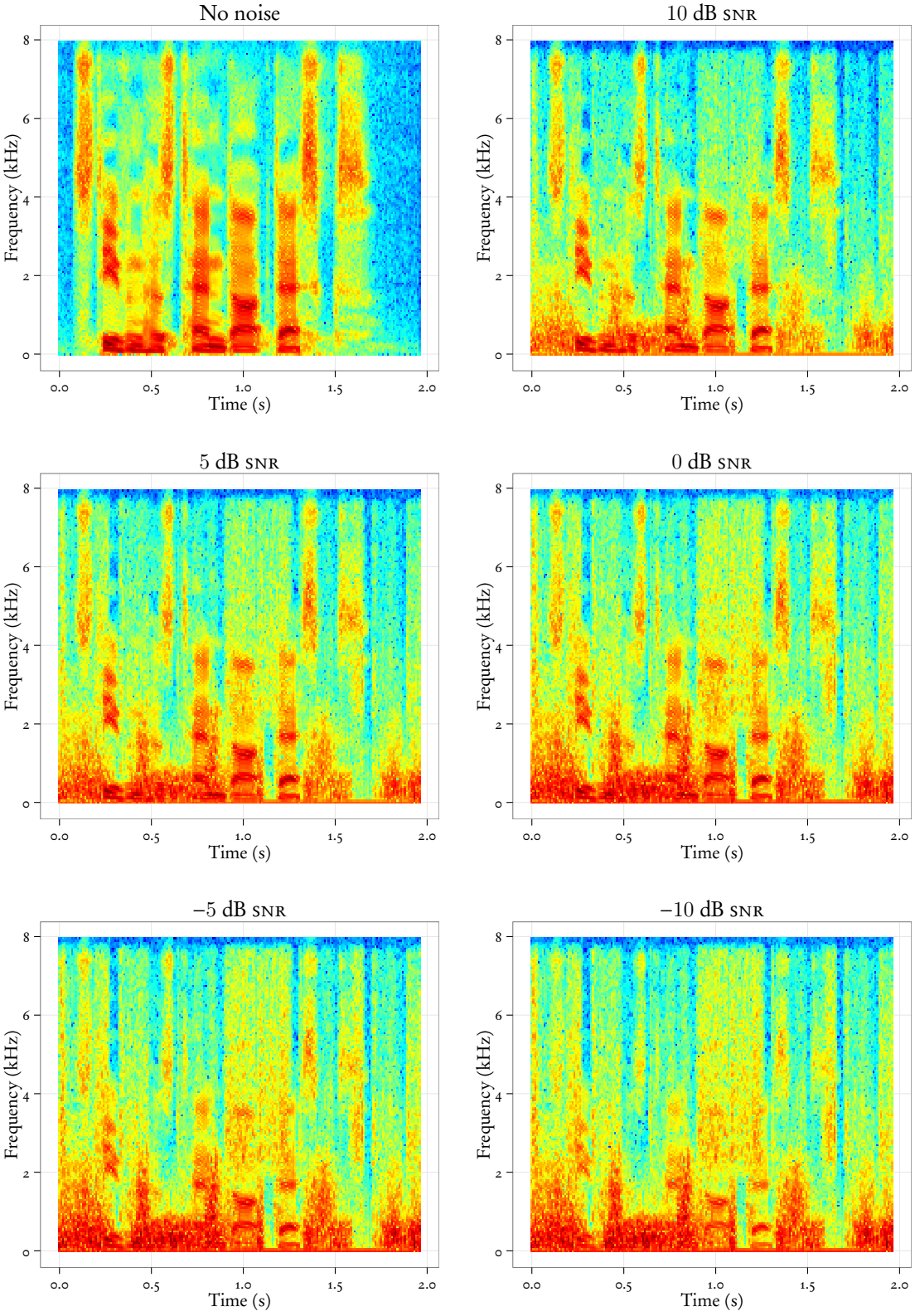

Figure 4.2: *Spectrograms of 'Steven has ten dark desks', without noise and with ICRA noise added at SNRs of 10, 5, 0, −5, and −10 dB*

adjustments to their speech in order to maintain its intelligibility to listeners. This modified speech is known as Lombard speech after the investigator who first documented the effect [Lombard, 1911]. The adjustments that can be made are extensive and include changes in amplitude, duration, pitch, formant frequencies, and short-term power spectra of vowels [Summers, 1988]. Similar techniques can be used to improve the intelligibility of synthetic speech (see, for example Cooke et al. [2013]).

Listeners, on the other hand, will try to maximize their reception of the speech by moving closer to the speaker, by cupping their ears, or by paying more attention to lip movements, body language, and the context of the spoken material. The effects of adjustments made by speakers and listeners on the intelligibility of speech at various SNRs and in the presence of a variety of noise types can be wide and varied and have been studied extensively by speech scientists for that reason.

The effect of the environment on intelligibility, other than the noise it contains, is largely due to the physical structure of the surrounding space, whether this be a park, a living room, or classroom. The multiplicative effect of sound bouncing off walls and other physical objects is known as reverberation and is dealt with in detail in Chapter 5.

4.2 Overview of experiments

Table 4.1 outlines the materials and methodologies used in the experiments carried out into synthetic speech in noise for this chapter, along with the number of the section in which the full details appear.

We wanted to evaluate the implementation of the experimental methodologies discussed in Chapter 3 whilst measuring the intelligibility of modern synthesis systems in ecologically-valid background noises. We piloted the implementation before carrying out a full experiment.

The full experiment indicated the presence of the crossover first noted in our baseline experiment, so we carried out a post hoc analysis of the *Blizzard Challenge* 2010 data, firstly, to check for any crossover and, secondly, to provide a measure of intelligibility with a system that had been tuned specifically for use in noise.

4.3 Ecologically-valid noises experiment

We had established a baseline for the performance of current speech synthesis systems from the experiment presented in Section 3.5. However, the results were limited by the fact that we had only used one specific background noise, namely that from track 5 of those validated by ICRA [Dreschler et al., 2001], and had not included natural speech.

Table 4.1: *Overview of synthetic-speech-in-noise experiments*

Experiment	Purpose	Systems	Stimuli	Noise	SNRs	Design	Participants	Sec.
Ecologically-valid noises	Can we use more ecologically-valid noises and what are their effects on synthetic speech?	*Roger*: HTS v. F-USEL	120 Matrix sentences	None, chat, music, both	−15 dB, −5 dB, and 10 dB	Mixed. Each heard one system with all combinations of noise and SNR	Native UK English: 36 in perception labs	4.3
Blizzard analysis	Does the crossover in intelligibility occur with other stimuli?	*rjs*: NS v. HTS v. F-USEL	78 SUSS 78 broadcast-news sentences	ICRA track 9	0 dB, −5 dB, and −10 dB	Mixed. Each heard all system/voice combinations in one SNR	Native English: 202 in perception labs	4.4

In this experiment, we sought to evaluate synthetic speech in more realistic noises and to include natural speech for comparison. In order to accomplish this, we had to select and deploy noises that were ecologically valid for the environments they would be used in and had to find alternatives to the ICRA noise as it was designed for use in laboratory settings and does not contain some common features of natural speech, such as: more than one speaker speaking at once; a female speaker; and the periods of quiet characteristic of natural speech. Moreover, it has its basis on read speech rather than conversational speech, which is known to have different prosodic properties and, particularly, fewer pauses [Howell and Kadi-Hanifi, 1991]. Additionally, of course, the ICRA noise does not account for the non-speech noises that might be found in the home environment.

Our studies using Amazon Mechanical Turk (AMT) in the previous chapter had shown conversation and radio/television (TV) as the most common background noises encountered by AMT participants. Since radio and TV often broadcast conversational speech and music, we chose to use noises that covered speech, music, and a combination of the two. We were particularly keen to ensure that our noises were as close as possible to those likely to occur in a real-life home environment whilst ensuring that participants were equally distracted by them. The conversation noise was, therefore, created by taking a sample from the Augmented Multiparty Interaction (AMI) corpus[1] [Carletta et al., 2006]. The sample consisted of four native UK English speakers, of mixed gender (as our listeners were expected to be), discussing the conceptual design of a remote control. Our motivation for choosing it was that the conversation was real with individual words being audible, but that the topic would be of no particular interest to our participants. The level of the speech varies, but we selected a section with minimal silence and took a long-term average

[1] http://corpus.amiproject.org

when calculating SNRs. Likewise, we chose a piece of music with minimal silence and instruments (violin and harpsichord) that would not be too familiar to our expected participants (undergraduate students). The music chosen was taken from the second track of the first compact disc (CD) from the collection, *Bach: The Six Sonatas for Violin & Harpsichord* [Laredo and Gould, 2007].

4.3.1 Method

The two base background noises (chat and music) were sourced as described above. The chat was selected by visually inspecting the waveform to identify a section without long silences (so that we would be able to achieve the desired SNR when added to speech) and listening for a good mix of speakers (to obviate the effect of gender). In this way, we kept the noise as realistic as possible whilst maintaining viability. The music was also chosen by visual inspection of the waveform, again by avoiding long silences and by listening for a mix of instruments. The noise samples were down sampled to match the speech and, in the case of the music, reduced from two channels to one. The background noise with both chat and music was created by combining the two base noises.

We wanted to be able to compare the performance of synthetic speech directly with natural speech (NS) and, therefore, had to choose a voice for which we could readily record the natural speech stimuli. For this reason, we used the RP-English voice *Roger* built for the 2010 variant of the HMM-based Speech Synthesis System (HTS), which we simply refer to as 'HTS' and the *Festival Multisyn* unit-selection engine [Clark et al., 2007], referred to as *Festival* unit-selection (F-USEL), thus ensuring that only the synthesis method varied, not the underlying speech data. The NS versions of the sentences were recorded with the original speaker, using the same equipment used for the recordings used to build the HTS and F-USEL versions. Generation of the synthetic speech stimuli did not include data from the recordings of the natural speech stimuli.

Matrix sentences were used as the basis of the speech stimuli. Seventy-eight Matrix sentences recorded for a previous study [Wolters et al., 2014] were supplemented with forty-two randomly generated. Each sentence was unique and was allocated to a batch of ten such that the repetition of words within the batch was minimized, although not eliminated entirely. The sound pressure level (SPL) of the speech and individual background noises was adjusted to 65 dB. We presented the stimuli at SNRs of −15, −5, and 5 dB, by reducing the SPL of the speech stimuli appropriately using the active speech level as described in Section 3.5.1. The lowest SNR is lower than the value at which the AMT experiment (Section 3.6) and previous work [Rhebergen and Versfeld, 2005] suggest fifty per cent intelligibility would be achieved. The value of

−5 dB represents the level at which our baseline experiment (Section 3.5) suggests that F-USEL and HTS perform equally and 5 dB is a value at which we would expect HTS to outperform F-USEL and ceiling effects to begin to appear.

For this experiment we used a mixed design, with the synthesis system being between-participants and background noise and SNR being within-participants. The choice of design made the number of conditions manageable and meant that any potential effect of hearing natural speech stimuli with synthetic speech was neutralized. The total number of conditions was thirty-six, derived from: three systems (HTS, F-USEL, and NS) ×four background noises (none, music, chat, and both) ×three SNRs (−15, −5, and 5 dB). All participants heard the same 120 sentences in one of the three systems, spread across the four background noise conditions and the three SNRs, such that each combination of background noise and SNR was presented in a batch of ten sentences. The distribution of sentences across all conditions was balanced using a Latin square.

Our previous experiments, like previous studies, had relied on participants wearing headphones to listen to stimuli, but we were aware that reminders might eventually be played through speakers and that some AMT participants might report using headphones when, in fact, they were using speakers and that this could undermine our results. For this reason, we decided to use speakers for our first experiment with noise so that we could determine whether this had any effect on the results. Participants listened to four blocks of thirty sentences from a pair of Genelec 8020A speakers (chosen for their very low harmonic distortion) located 1 m in front of them and 1 m apart. Each of the four blocks was accompanied by one of the background noises played through another pair of the same speakers located 1 m behind the participant and 1 m apart. The output of the speakers was equalized by taking sound level meter readings at the midpoint of the 'X' described by an imaginary line connecting the centres of the diagonally opposite pairs of speakers. The condition with no background noise was always presented in the first block, but the order of the other three was determined by a Latin square. The background noise, when present, was started automatically at the commencement of the block (before the presentation of the first stimulus) and played continuously for its duration. The distribution of SNRs within a block was randomized.

The experiment was carried out in a studio in the sound laboratories at the Centre for Speech Technology Research (CSTR). The room was chosen as being the closest to the average English living room [Department for Communities and Local Government, 2012], although it has a smaller floor area (approximately 13 m^2 versus approximately 17 m^2) and a slightly greater height (approximately 3 m versus 2.5 m) and for having the space to have speakers in front of, and behind, the listener. The

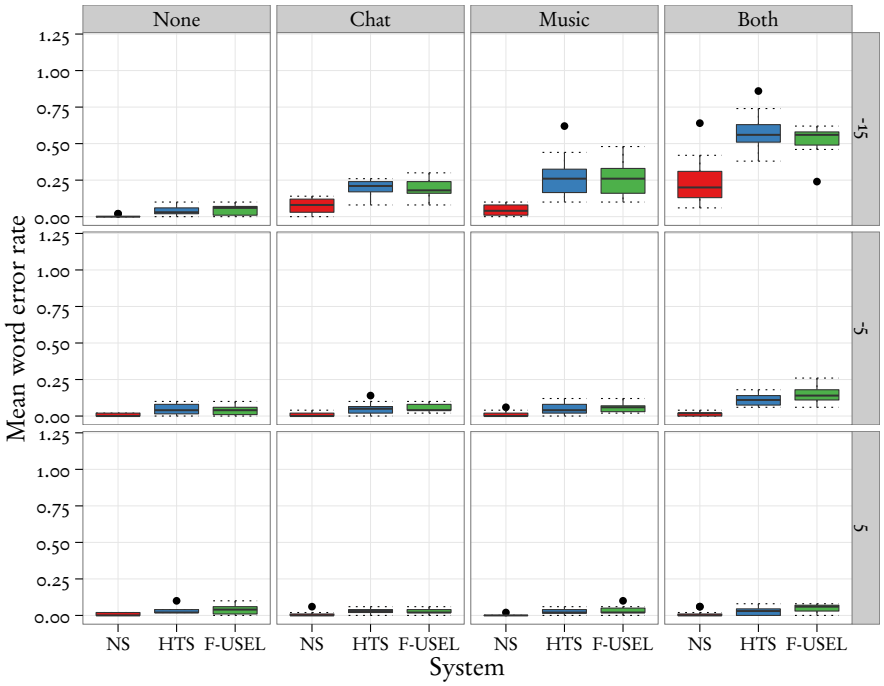

Figure 4.3: *Mean WER by system, background noise, and SNR (dB) of pilot experiment*

studio is sound treated, so that any extraneous background noise or reverberation would have been eliminated.

Our intention was to compare our results with those from our baseline experiment. Specifically, we wanted to know: whether fifty per cent intelligibility is reached at −15 dB SNR; the relative performance of HTS and F-USEL at −5 and 5 dB SNRs; whether there was any difference in performance when using speakers rather than headphones; and in what noise the crossover we had previously observed occurred, if at all. Since this experiment used the 2010 variant of HTS rather than the 2007, and speakers rather than headphones, used in previous experiments, it was first piloted and the mean word error rates (WERS) for each of the systems, broken down by background noise and SNR, plotted as in Figure 4.3. We can clearly see the pattern that would be expected, that is, increased WERS for synthetic over natural speech, for the lower SNRS, and for the mixed background noise. The structure of the results suggested that the design of the experiment was sound, so the experiment proper was run with a total of 36 native UK English speakers, 15 male and the remainder female. As for previous experiments, participants provided the demographic and other data outlined in Table 3.2 on page 41 before typing in what they heard of the speech stimuli into a web browser. Participants heard, and transcribed, six practice sentences before hearing the main stimuli.

Participants were recruited through the University of Edinburgh's student recruitment service and reimbursed £5 for their time. Their profile was typical of a student

cohort with all 36 of them being aged between 18 and 29; 35 being full-time students; and 44 % having completed high school, 39 % a college education or a bachelor's degree, and 17 % a master's degree. No participant wore a hearing aid. However, 2 of the participants scored 10 or higher on the Hearing Handicap Inventory for Adults Screening Version (HHIA-S) and a further 2, who scored low on the HHIA-S, mentioned hearing problems in the free comments, although none of them had exceptionally high mean WERs. No significant effects ($R^2 = 0.17$, $F(8, 27) = 0.7$, $p = 0.689$) were found from an analysis of variance (ANOVA) of a linear model built to predict the mean WER from potential hearing loss (an HHIA-S score of ten or higher), a reported problem with hearing, synthetic speech listening frequency, age group, gender, and whether participants described themselves as computer scientists.

4.3.2 Results

The results of the full experiment, shown in Figure 4.4, bear out the results of the pilot, albeit with more variance. The results are generally as one would expect, with WER increasing in more difficult listening conditions. However, it does appear that listening in the presence of music was more detrimental to intelligibility than listening in chat; certainly at the −15 dB SNR. This is contrary to what might be expected given the expectation (elucidated in Section 4.1, based on Figure 4.1) that increased similarity of a noise with a signal causes increased disruption to its intelligibility. However, it might be that the music noise offers the listener fewer glimpses of the speech stimulus, thus making it harder to perceive [Cooke, 2003]. Inspection of the spectrograms of the noises given in Figure 4.6 on page 92 shows that this is likely to be the case.

In order to visualize the performance of the synthesis systems against each other and the natural speech, we plotted the mean number of errors for each background noise and each SNR, as shown in Figure 4.5. Taking the chat noise first, as it is the closest to that used in our baseline experiment, we can see that the performance of HTS and F-USEL is comparable to that of the baseline experiment shown in Figure 3.1 on page 46. However, the errors at −15 dB do not reach 2.5 (that is, 50 % intelligibility) as we might have expected and, indeed, is the case for the conditions with music and both chat and music. Moreover, the WERs are lower across the board. At −5 dB, in all conditions except music, the performance of HTS and F-USEL is effectively equal, with equality occurring between −5 and −10 dB for the music condition. At the SNR of 5 dB, HTS performs as well as, or better than, F-USEL in all conditions, although it has to be said that the difference between them is small.

If the background noises we had chosen were to produce equivalent results to

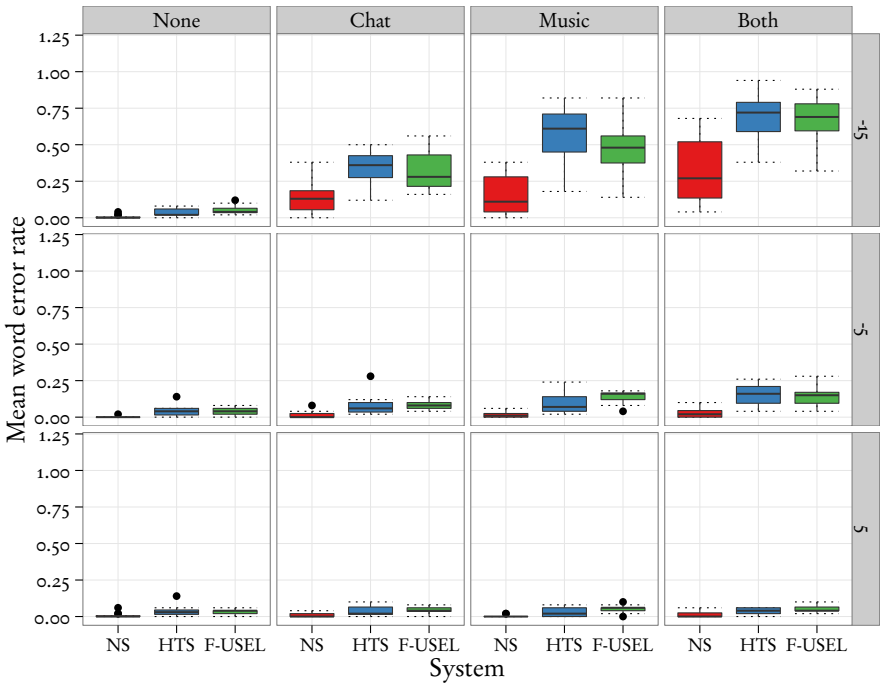

Figure 4.4: *Mean WER by system, background noise, and SNR (dB) of full experiment*

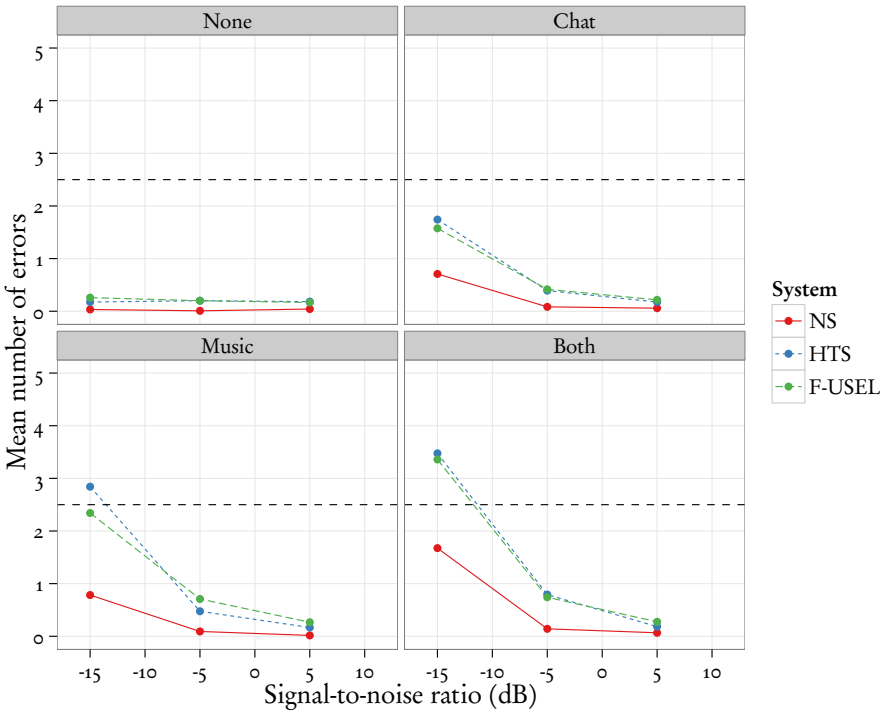

Figure 4.5: *Mean number of errors by system, background noise, and SNR (for the condition without noise, figures for SNR represent how much the intensity of the speech was reduced)*

those using the ICRA noise in the baseline experiment, we would expect to see the same crossover in intelligibility between HTS and F-USEL that we had seen previously. Figure 4.5 clearly shows that the crossover is present. However, it does seem to vary in somewhat unexpected ways depending on the background noise. In the chat noise, the crossover is as we saw it in the baseline experiment with the ICRA noise; in music, the crossover is much more pronounced; but in music and chat combined, the crossover lessens. Perhaps most surprisingly, there is a crossover in the *opposite* direction when the speech stimuli are in quiet, that is, the levels of the speech stimuli were simply reduced to achieve the level of speech equivalent to that in the noise conditions.

As for previous experiments, we used generalized linear mixed models (GLMMS) to analyse the data further, this time using the formula given in Equation (4.2).

$$\text{Error} \sim \text{System} * \text{Music} * \text{Chat} * \text{SNR} + (1 \mid \text{Sentence}) + (1 \mid \text{Participant}) \quad (4.2)$$

We chose the individual-level predictors System, Chat (whether present or not), Music (whether present or not), SNR, and all their interactions and modelled their effect on the number of errors participants made. The group-level predictors were the sentence and participant identifiers, for both of which we allowed a random intercept. We removed each of the predictors (and its interactions) in turn from the fully-specified model and carried out an ANOVA between the full and reduced models. The results are shown in Table 4.2, from which it can be seen that each of the predictors alone was found to be significant at the $p < 0.001$ level. As we had found in previous experiments, the interaction System × SNR was also significant ($p < 0.001$). Of the interactions accounting for the crossovers in Figure 4.5, only System × Music × SNR is significant ($p < 0.001$).

4.3.3 Discussion

The purpose of this experiment was to build on the baseline experiment, using the experimental methodologies we had evaluated in Chapter 3, by establishing the performance of modern synthesis systems in the presence of noise that was more realistic and ecologically-valid than noises used in previous studies. The experiment was a success in that we obtained meaningful results with two new noises, individually and in combination.

In general, we found that the results from this experiment support those of the baseline experiment presented in Section 3.5.2 and demonstrate that, despite recent advances, synthetic speech intelligibility suffers significant degradation in the presence of noise. The use of the 2010 variant of HTS in this experiment rather that the 2007 used previously, demonstrates that the effect has persisted over developments in

Table 4.2: *Significance of individual-level predictors (ANOVA model comparison, χ^2 test) for the GLMM of the ecologically-valid noises experiment*

Predictor	AIC	*p*
Baseline	7929	—
System	7968	0.001
Music	8256	0.001
Chat	8130	0.001
SNR	8849	0.001
System x Music	7929	1.000
System x Chat	7929	1.000
Music x Chat	7929	1.000
System x Music x Chat	7926	0.519
System x SNR	7929	0.001
Music x SNR	7929	0.001
System x Music x SNR	7929	0.001
Chat x SNR	7929	1.000
System x Chat x SNR	7929	1.000
Music x Chat x SNR	7929	0.001
System x Music x Chat x SNR	7929	0.114

HTS. More specifically, the inclusion of natural speech enabled us to show that it is still significantly more robust to reductions in the SNR than synthetic speech despite the intervening years of research since any major studies into synthetic-speech intelligibility in noise were carried out. The robustness is evident from the shallower slopes of the lines representing NS in Figure 4.5.

Furthermore, the different results obtained from using the music noise, in isolation and combined with chat, demonstrate the need to test synthesis systems in a range of the noises in which they are likely to be deployed, if a full understanding of their potential performance is to be achieved. For example, the performance of both systems with the music noise was better than might have been predicted from their performance in chat (based on the assumption that background speech is likely to cause the most interference to speech) and the much more pronounced crossover suggests considerable thought should be given to the range of SNRs that will be encountered when choosing a synthesis system. To visualize the effect that our background noises might have on speech, we created the spectrograms in Figure 4.6. The spectrograms represent the first 10 s of the chat, music, and both noises used in the experiment. For comparison purposes, we have also included the ICRA noise used in our baseline experiment. It is clear that the chat noise contains some periods of silence, which allow listeners larger glimpses of the stimuli and, therefore, to reduce their WER. In contrast, the music and both noises having a more even spread of energy across time and frequency, thus accounting for the higher WERs in these noises. The energy in the ICRA noise, although not as evenly distributed as in the music and both

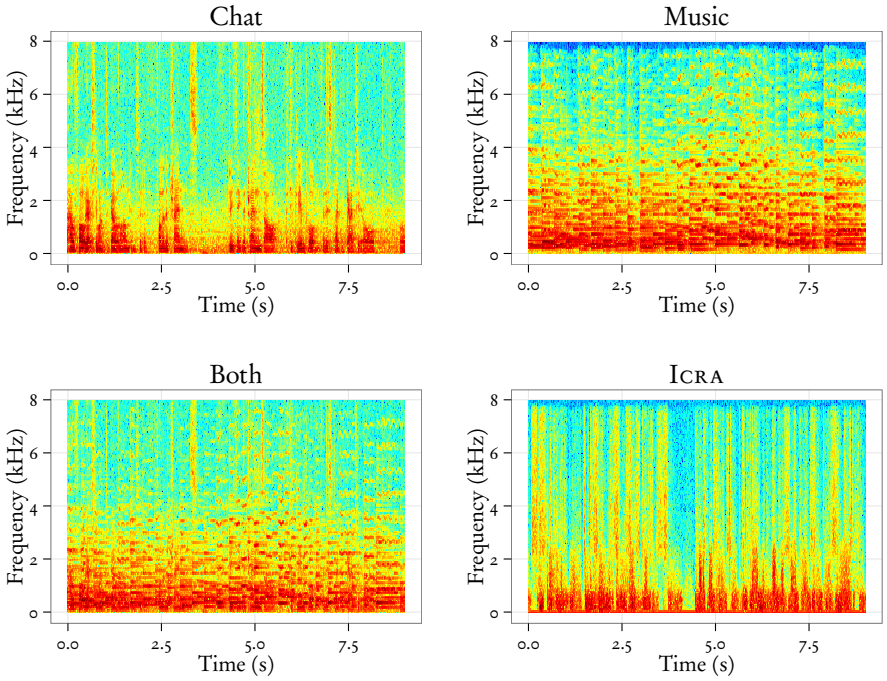

Figure 4.6: *Spectrograms of the noises used in the ecologically-valid noises experiment (chat, music, both) and the ICRA noise used in the baseline experiment*

noises, does not fluctuate as randomly as the chat noise, which probably explains the higher WERs associated with it.

A comparison of the relative results from this experiment with the baseline shows that whether participants listen to our stimuli through speakers or headphones appears to have no appreciable effect on the results. This conclusion was confirmed by a contemporaneous, but more rigorous, treatment of the issue [Raitio et al., 2012], which found the same relative differences between systems with speakers and mono or stereo headphones.

Our results also lend credence to our supposition that the crossover found in the baseline experiment was not an experimental artefact, but a real effect of the interaction between speech synthesis system and SNR. Even though the ICRA noise was replaced with one of three background noises, all of them—to a greater or lesser degree—reproduced the crossover. It seems clear from the persistence of the crossover, its variability against different background noises, and the GLMM, that current state-of-the-art systems fare differently in varying levels and types of background noise. This highlights both the importance of selecting a synthesis system to match the expected environment and the consequent importance of furthering our understanding of synthetic speech intelligibility through research of this kind. It is not clear why mixing music and chat should result in a less marked crossover.

We were intrigued by the recurrence of the crossover between HTS and F-USEL in the noisy conditions and sought to investigate this further. Our concern was that it

might be caused by the use of Matrix sentences and wanted to rule out this possibility. We, therefore, chose to analyse the data from the most recent *Blizzard Challenge* as this had included speech in noise with two different types of sentence, neither of which were Matrix sentences.

4.4 Analysis of the *Blizzard Challenge* 2010 data

After the results of the baseline experiment had been submitted for publication, we were approached by the organizers of the *Blizzard Challenge*, who wanted to use our experiment as the basis of a section of the *Blizzard Challenge* 2010 involving speech in noise [King and Karaiskos, 2010]. In fact, there were two sections: ES2 for English and MS2 for Mandarin. We provided the scripts to create the stimuli for the challenge and, in return, were able to use the results to broaden our investigation of synthetic speech in noise by including broadcast speech and semantically unpredictable sentence (SUS) stimuli.

We wanted to know whether the results from the *Blizzard Challenge* reflected those from our baseline experiment. We hypothesized that the relative results would be the same and that we would see the same crossover that was first seen in the baseline experiment and had persisted into the ecologically-valid noises experiment.

The ultimate goal of our work is to contribute to the improvement of the intelligibility of synthetic in noise, so we were particularly interested in seeing whether, and by how much, the intelligibility of synthetic speech could be improved when it was specifically tuned to be used in a noisy background as the systems in this challenge had been.

4.4.1 Method

Researchers and commercial providers of speech synthesis systems were invited to make submissions for one or both of the challenges ES2 and MS2 of the *Blizzard Challenge* 2010. For ES2, submitters were asked to build a system from Phonetic Arts' *rjs* speaker specifically for the assessment of intelligibility in additive noise, with no consideration of naturalness or speaker similarity. No further information on the type or severity of noise was given prior to submission. With this system, submitters were asked to synthesize both broadcast-news and semantically unpredictable sentences. The same criteria were given for MS2, except that the system was to be built from a Mandarin corpus from the National Laboratory of Pattern Recognition, Institute of Automation of Chinese Academy of Sciences. Since the study of languages other than English is outside the scope of this current work, our focus here is on challenge ES2

only and no analysis of MS2 was undertaken.

After the synthesized sentences were received by the challenge organizers, noise was added at SNRs of 0, −5, and −10 dB using the procedure described in Section 3.5.1 for our baseline experiment, except that track 9 from the ICRA CD was used as the noise. Track 9 differs substantially from track 5 and consists of 3-band speech modulated noise from three male and three female speakers with an idealized speech spectrum and speaking effort raised against a perceived background noise. Also included in the challenge were three reference systems: A, B, and C. System A was the natural speech of the speaker who recorded the corpus used for the challenge; System B was the *Multisyn Festival* unit-selection system [Clark et al., 2006]; and System C was a speaker-dependent hidden Markov model (HMM)-based system with automatically produced labels built for the 2005 *Blizzard Challenge* [Zen and Toda, 2005] (an earlier version of the HTS system used in our baseline experiment). Since the exact systems used for our baseline experiment were not entered into the challenge, we shall use these systems for comparison and refer to them as NS, F-USEL, and HTS, respectively.

Participants who listened to the stimuli were either: volunteers or experts from submitting teams and other sources, who listened over the Internet with their own equipment; or were paid native English-speaking students recruited in Edinburgh, who listened in sound-treated listening booths. Sentences, stimulus type (broadcast or SUS), systems, and SNRs were balanced so that no participant heard the same sentence twice and heard one example from each condition.

For the analysis presented below, we followed the protocol adopted for earlier laboratory experiments and excluded all participants who had not been recruited for the listening booths, had declared themselves as non-native speakers, or had not attempted the full complement of sentences for the group they were allocated to. Even with these restrictions, we were left with 202 lab participants—a far larger cohort than we could have recruited.

4.4.2 Results

Two types of stimuli had been used for the challenge: broadcast-news and semantically unpredictable sentences. Figure 4.7 shows the performance of NS, F-USEL and HTS with both sets of sentences.

Comparing with our baseline experiment, broadcast-news sentences are probably the closest to the Matrix sentences we used, in that they make semantic sense, although they differ by having a variable length and format and do not have a fixed vocabulary. This similarity is reflected in the similarity of the results obtained from

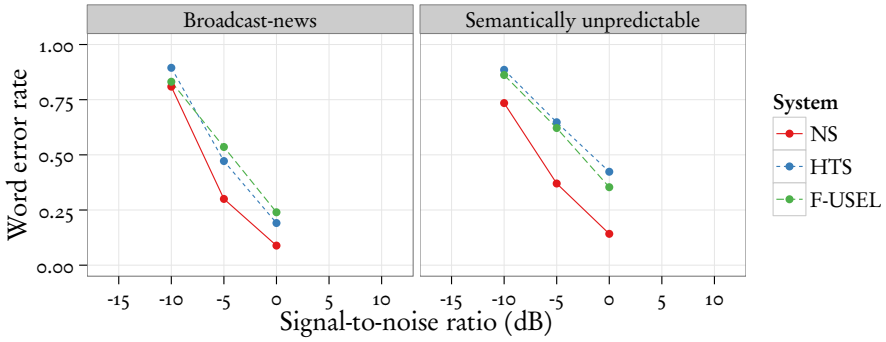

Figure 4.7: *WER by system, stimulus type, and SNR for NS, HTS, and F-USEL*

both experiments, with the crossover in performance being clearly visible at about the same SNR (compare Figure 3.1 on page 46 and the left section of Figure 4.7). However, the WER is considerably higher, probably partly because the ICRA noise selected for the *Blizzard Challenge* is more challenging and partly because the broadcast-news sentence material is less predictable and has a higher preponderance of uncommon words.

In the case of the ecologically-valid noises experiment, which also used Matrix sentences rather than broadcast-news, the closest noise condition to the ICRA noise used here is 'Chat', the results of which are depicted in the top-right section of Figure 4.5 on page 89. Once again, the pattern of results is very similar to those in the left section of Figure 4.7, albeit the crossover appears more pronounced in the *Blizzard Challenge* data and the WERS are much higher, probably for the reasons already given.

Interestingly, the right-hand section of Figure 4.7 shows no crossover between the performance of HTS and F-USEL when the stimuli are SUSS and HTS performs slightly worse than F-USEL across all SNRS. It seems from the graphs that NS and F-USEL perform almost identically with both types of sentence, but that the performance of HTS is more variable, becoming more stable with SUSS. We cannot be sure why this might be, but speculate on it being the result of tuning and note it for further investigation in Chapter 7.

Our secondary reason for analysing the data from the *Blizzard Challenge* was to examine how intelligible it is possible for synthetic speech to be when it has been tuned specifically for use in background noise. As Figure 4.8 shows, one of the systems, System N, was able to outperform NS in both conditions where SNR was negative, that is, the noise was louder than the speech. This was possible even when the nature of the background noise was not known in advance and shows, firstly, how much progress can be made in improving the intelligibility of synthetic speech and, secondly, how important it is to determine the nature of domestic background noise

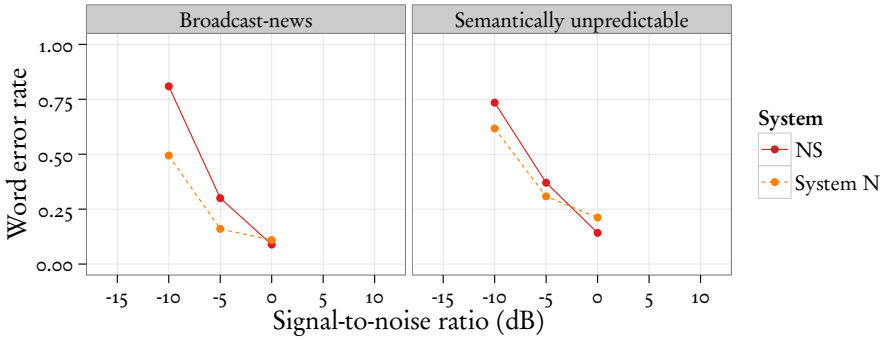

Figure 4.8: *WER by system, stimulus type, and SNR for System N and natural speech*

and the effect it has on intelligibility.

4.4.3 Discussion

The crossover in the performance of HTS and F-USEL first seen in our baseline experiment and repeated in the ecologically-valid noises experiment is also present in the *Blizzard Challenge* data. Therefore, it has persisted across three versions of HTS and five different background noises (albeit some of them were very similar). The results obtained from the analysis of the *Blizzard Challenge* data in which broadcast-news sentences were used, suggest that the crossover is not just a side effect of using Matrix sentences. However, it is still possible that there is some correlation with whether the stimuli used are semantically predictable, as the crossover is never seen with SUSS.

In addition to wanting to check for any crossover in the *Blizzard Challenge* data, we had also wanted to compare the results of the reference systems with those from our ecologically-valid noises experiment. Figure 4.9 shows mean WERS by system and SNR for each of the three reference systems with broadcast-news and semantically unpredictable sentence as stimuli. We can compare the results for −5 dB SNR with the chat noise at the same SNR in Figure 4.4 on page 89. Bearing in mind that the chat noise yielded lower absolute WERS than the ICRA noise, the relative performance of the three systems is comparable across the two experiments.

4.5 Conclusion

Background noise is a common, and potentially fatal, problem for spoken communication and, therefore, its effects on natural speech have been widely studied. Unfortunately, its effects on synthetic speech, although usually more destructive, have not received the same level of attention and the use of noise in speech synthesis research has often been limited to the use of static maskers as a means of increasing discrimination in intelligibility tests.

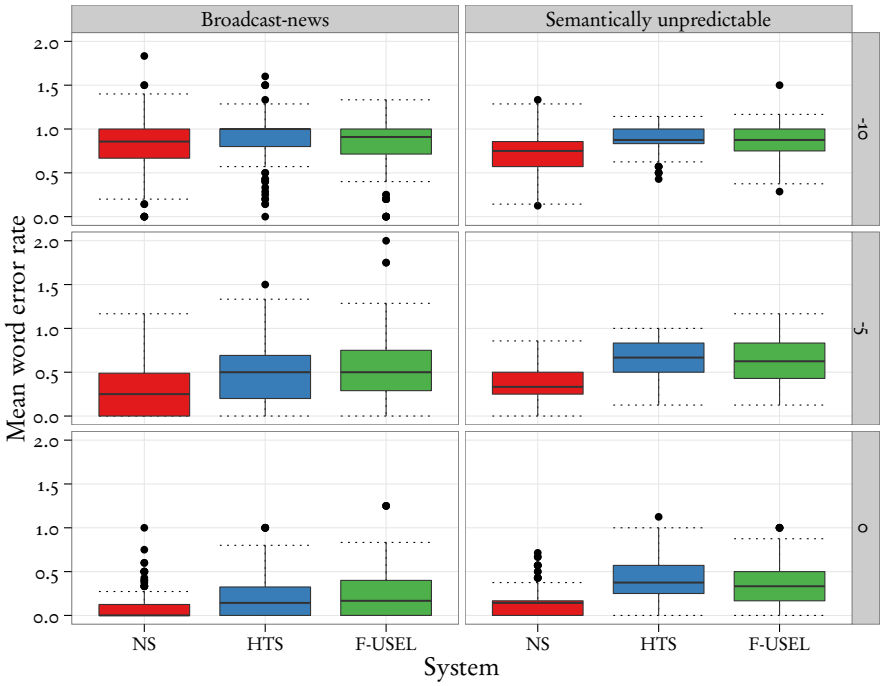

Figure 4.9: *Mean* WER *by system and* SNR *with broadcast-news and semantically unpredictable sentences for the* Blizzard Challenge

Our need to provide synthetic speech for potentially noisy environments necessitated the development of realistic fluctuating noises for testing purposes. We used the results from our experiments on AMT to identify the most common types of noise and develop ecologically-valid noises. We put our noises to the test in a laboratory experiment and carried out an analysis of the *Blizzard Challenge* 2010 data, the stimuli for which we had helped to generate.

We successfully trialled our new ecologically-valid noises and achieved similar results, in relative terms, to those from the baseline experiment and to the *Blizzard Challenge* 2010. This leads us to believe that the assessment of synthesis systems relative to one another, and natural speech, can be achieved using our noises. Moreover, it seems that similar results are returned whether participants listen through headphones or speakers. However, absolute WERs from the ecologically-valid noises experiment were lower than for any of the others.

The crossover we had first noticed in the baseline experiment recurred in subsequent experiments, but not in the *Blizzard Challenge* results attained from using SUSS. We suspect that this might be evidence of SUSS having an unnatural effect on intelligibility.

CHAPTER 5

SYNTHETIC SPEECH IN REVERBERATION

NOISE is not the only source of interference to act on human speech. The sound waves of speech decay over time as they travel through a space and are reflected back from objects in that space. These decaying sound waves are known as reverberation and result in the sound from the first part of a sentence still being present when a later part is uttered. How much interference is caused by this phenomenon depends on a number of factors, including: the 'strength' of the reverberation, the frequencies of the speech, the placement of objects in the space, distance from the speech source, and the perceptual abilities of the listener.

Reverberation plays an important part in the perception of any sound by humans and its significance to the perception of speech has been known for some time. Consequently, it has been the subject of much research. Unfortunately, most of the research has been focused on specific scenarios, such as in the classroom, in the auditorium, or with hearing aids and has been orthogonal to synthetic speech research.

5.1 The problem of reverberation

The arrival of sound waves at the human ear can be classified into direct, early, and late. Direct signals are those that arrive from the source without reflection whilst early reflections are those signals that have been reflected, but arrive within about 0.05 s, and late arrivals are those reflections that arrive after more than about 0.1 s.

Generally speaking, direct signals give us the best representation of the sound; whilst early reflections are used as clues to the size and shape of the environment and help us locate the sound; and late arrivals form what we normally think of as reverberation and are detrimental to our perception of sound.

The rate at which reflections accumulate in a space is proportional to the square root of its volume. This, in turn, means that signals eventually lose sufficient energy that they are no longer perceptible to the human ear. The time taken for a signal to decay is known as its reverberation time (RT) and the standard measure of potential reverberation in a room is the time it takes for a signal to reduce in intensity by 60 dB, known as its RT_{60}.

The RT_{60} of a room is calculated using the following formula derived empirically by Sabine [1923]

$$RT_{60} = \frac{24\ln(10)V}{c\sum_{i=1}^{n}(s_i a_i)} \tag{5.1}$$

where V is the volume of the room in m^3, c is the speed of sound at 20 °C, and s_i and a_i represent the surface areas and noise absorbency of all the surfaces in the room. Therefore, a small room with noise-absorbent wall coverings will have a smaller RT_{60} than a cathedral with bare stone walls and, consequently, much lower potential for reverberation.

The situation for speech is somewhat more complicated, not least because different frequencies within a signal will have different RTs and, as we saw in Chapter 4, the different components of speech occur at different frequencies within ranges that vary according to age, gender, and even between individuals. In general, high frequencies decay more quickly than low, so that in reverberant conditions vowels sound fuller and tend to mask consonants, making it difficult, for example, to distinguish between the words cat, cad, and can.

However, some reverberation is generally considered beneficial to speech perception, making it sound clearer and sharper, although too much late reverberation can make speech incomprehensible. Boothroyd [2005] gave the visual analogy of text, saying that early reverberation could work like a sentence printed twice slightly out of phase, making it stand out; whereas late reverberation had the effect of a sentence printed many times out of phase, making it disappear. The result is exemplified by the two sentences below.

The effect of early reverberation. 

In the case of natural speech, Bistafa and Bradley [2000] found that 100 % intelligibility could still be achieved at an RT_{60} of 0.4 s to 0.5 s.

The effect of objects in a room on reverberation will be well known to anyone who has emptied a room for decorating or has replaced a carpet with a solid floor. What may not be quite so clear is that the *position* of objects in a space can have a significant effect, since the position affects how sound 'flows' around and between them. A good visual analogy would be to imagine the room filled with water and how the ripples would look if a pebble were dropped in at the same point as the sound source. Sound waves produce a similar result—sometimes acting together to make a bigger wave and sometimes crashing into one another to make a 'choppy' area.

The analogy of ripples in water is also useful when considering how the effect of reverberation will differ depending on its distance from the source. Since speech reception will be a combination of the speech received directly and that reverberated, the relative levels of each will depend on the listener's distance from the speaker. Clearly, there will be a point at which the two levels are equal and this is known as the critical distance. The critical distance can be calculated as

$$d_c = \frac{1}{4}\sqrt{\frac{\gamma \sum_{i=1}^{n} (s_i a_i)}{\pi}} \tag{5.2}$$

where γ is the directionality of the signal from 0 to 1 and s_i and a_i represent the surface areas and noise absorption of all the surfaces in the room. If the volume of the room is V (in m^3) and the RT_{60} has already been calculated from Sabine's formula in Equation (5.1), the critical distance can be calculated using

$$d_c = 0.057\sqrt{\frac{\gamma V}{RT_{60}}} \tag{5.3}$$

In the same way that they do for speech in noise, individuals differ in their ability to perceive speech in reverberation, not only because of differences in their physiology, but also because of differing mental acuity. For example, people with a hearing impairment will suffer the same relative degradation of intelligibility as those with normal hearing, but at a lower absolute level [Payton et al., 1994] and musical training of more than eight years' duration is known to make participants more resilient to reverberation than those with no more than three years' training [Bidelman and Krishnan, 2010].

When carrying out experiments involving reverberation, it is common to capture the acoustic characteristics of a room by recording its impulse response. The impulse response is how the room responds to a very short signal (impulse). Since the way a room responds and produces reverberation is essentially a mathematical function, the capture of an impulse response effectively involves recording a known signal with its

Table 5.1: *Overview of synthetic-speech-in-reverberation experiments*

Experiment	Purpose	Systems	Stimuli	Reverb	Design	Participants	Sec.
Low-level reverberation	What is the effect of low reverb on synthetic speech? Does AMT match the lab?	*Roger*: NS v. HTS v. F-USEL	90 Matrix sentences	None Lounge: 100 cm and 250 cm	Within-subjects. Each heard all systems in all reverbs	Native UK English: 36 in perception labs Native US English: 65 on AMT	5.3
High-level reverberation	What is the effect of high reverb?	*Roger*: NS v. HTS v. F-USEL	90 Matrix sentences	None Lounge: 250 cm and Aula Carolina	Within-subjects. Each heard all systems in all reverbs	Native US English: 35 on AMT	5.4

reverberation and then removing the signal. We are then left with a recording that can be combined with a stimulus (in our case speech) through a process known as convolution to achieve a very similar reverberation to that which would have resulted from playing the stimulus in the room, whilst excluding the unwanted effects of using a real room. The convolution operation is summarized by Loizou [2007] as Equation (5.4), where x is the vector to be convolved and h is the vector containing the impulse response.

$$y(n) = \sum_{-\infty}^{\infty} x(k)h(n-k) \tag{5.4}$$

The practical effect of reverberation can be seen from Figure 5.4 on page 113, which shows spectrograms of an utterance used as a stimulus for the experiments in this chapter. Without reverberation (top left), the speech is clearly demarcated, but as the level of reverberation increases, the speech appears 'smeared' to the right (later in time), clearly demonstrating how reverberated earlier speech interferes with later speech.

5.2 Overview of experiments

Having established in Chapter 4 that synthetic speech still suffers more than natural speech in the presence of noise and that different synthesis systems perform differently, we turned our attention to the effects of reverberation. We wanted to explore how reverberation affects synthetic speech and whether the results from the lab would be reflected on Amazon Mechanical Turk (AMT).

We carried out two experiments, the first in the lab and on AMT with a low level of reverberation and the second on AMT that included a much higher level of reverberation. Table 5.1 provides an overview of the two experiments.

5.3 Low-level reverberation experiment

Our main aim in this experiment was to put our experimental methodology to the test and establish the relationship—if any—between the intelligibility of natural speech and the two synthesis systems we had used in previous experiments in reverberant conditions.

Our secondary aim was to establish the viability of running experiments involving reverberation on AMT. We had concerns that the noise present in the environments AMT participants were using might interfere with the results of reverberation experiments.

We hypothesized that:

- overall, increasing levels of reverberation would interfere more with intelligibility
- because the HMM-based Speech Synthesis System (HTS) has a tendency to sound buzzy, particularly through headphones, a moderate level of reverberation would disguise this and, by reducing the distraction to the listener, make it more intelligible than with no reverberation
- AMT would provide the same relative rankings as the laboratory.

We were also interested in seeing whether the first two hypotheses would result in the crossover effect seen in the experiments with additive noise.

5.3.1 Method

In line with previous experiments, we carried out listening tests using Matrix sentences in three levels of reverberation. Since it would have been impractical to have participants attend a room with the desired acoustic characteristics and impossible for AMT participants to do so, we sought to create reverberant stimuli that could be played through headphones, but would simulate listening to the stimuli in a reverberant room. Using headphones confers the further advantage that each participant experiences exactly the same reverberation rather than one that has been altered by their position in the room or objects in the room having been moved.

We were particularly keen to find a binaural impulse response that we could convolve with our speech since there is known to be a significant difference between monaural and binaural speech perception in reverberation [Nabelek and Robinson, 1982]. We also wanted to use reverberation levels typically encountered by real users in real environments and, to this end, we were fortunate enough to be given permission to use the impulse responses recorded for the Computational Hearing in Multisource Environments (CHiME) project at the University of Sheffield [Christensen et al.,

2010]. The purpose of the CHiME project was to establish a framework for the assessment of automatic speech recognition (ASR) in the presence of reverberation, which necessitated the recording of impulse responses that could be used to add the effect of reverberation to speech files. Binaural impulse responses (BIRs) were recorded in a semi-detached, Victorian house, typical of those found in the UK. Recordings were made systematically, in two rooms, from varying distances and azimuths with responses to a sine sweep played from a B&K 4227 artificial mouth being recorded using a B&K head and torso simulator (HATS) at the same location.

The BIRs used for this experiment were those recorded in the lounge, a room measuring 385 cm by 385 cm and 365 cm high—a 13 % smaller floor area than the UK average and nearly 50 % higher ceiling [Department for Communities and Local Government, 2012]. The room was measured as having a reverberation time (RT_{60}) of 300 ms using Schroeder integration. Using Equation (5.3), the maximum critical distance (that is, the point beyond which the reverberant, rather than the direct, sound field is predominant) for the room was calculated to be 0.765 m. We used the impulse responses from recordings at 100 cm and 250 cm distance and 0° azimuth (i.e. head-on). Clean synthetic speech was normalized to 65 dB sound pressure level (SPL) before being upsampled to 96 kHz and convolved with each channel of the appropriate impulse response. To simulate a condition without reverberation, an equal length file of a one and all zeroes was used as the impulse response. After convolution, the resultant 'tail' was removed and the mixed signal downsampled to 16 kHz.

The use of a stereo signal meant that, unlike in previous experiments, the stimuli had two channels rather than one. The result is that participants' left and right ears would hear slightly different stimuli, which introduced the possibility of their handedness having an effect on the results. We, therefore, added the Edinburgh Handedness Inventory [Oldfield, 1971] to our pre-experiment questions to measure how left- or right-handed participants were so that we could check for any effect.

Ninety Matrix sentences in natural speech (NS), the 2010 variant of the HMM-based Speech Synthesis System (HTS), and *Festival* unit-selection (F-USEL) were taken from those generated as described for our ecologically-valid noises experiment in Section 4.3.1. Each sentence was heard only once by each participant, who heard ten sentences in each of nine conditions (NS, HTS, and F-USEL speech without reverberation ('none') and with the level of reverberation that would have been experienced from distances of 100 cm and 250 cm). The two distances were chosen to be as near as possible to the critical distance and at a point well beyond it. The presentation of conditions, and sentences across conditions, was balanced using a Latin-square design and participants were allocated to a group at random.

Listeners were recruited either from the University of Edinburgh's recruitment service, were compensated for their time with £8; or through AMT and compensated with US$1. Lab participants listened to the stimuli with headphones in a listening booth at the Centre for Speech Technology Research (CSTR) as described in Section 3.5.1. The participants from AMT listened in their own environment with their own equipment, but were instructed to wear headphones at all times and to record what features they had.

Lab participants consisted of a total of 26 females and 10 males, all of whom were native speakers of UK English. They were overwhelmingly full-time students (34 full-time students, 1 employed full-time; and 1 employed part-time) and followed a typically student profile in terms of age and education, with 33 % being under 20; 64 % 20 to 29; and 3 % 40 to 49; 3 % not having completed high school; 53 % having completed high school; 19 % a bachelor's degree; and 25 % some college. Age groups and genders were distributed evenly across groups (age: Fisher's Exact test, $p = 0.528$; gender: Fisher's Exact test, $p = 0.875$).

Surprisingly, nearly 39 % of participants scored 10 or higher on the Hearing Handicap Inventory for Adults Screening Version (HHIA-S), suggesting that they may have sensorineural hearing loss. A further 7 reported minor hearing problems, but none was fitted with a hearing aid. None of those who reported problems or had a high HHIA-S score had remarkable word error rates (WERS) and a linear model predicting an appropriate transformation (power of −0.2) of the mean WER did not show any significant effect. Nor did the same model show any significant effect of synthetic speech listening frequency, age group, gender, or whether participants described themselves as computer scientists ($R^2 = 0.27$, $F(9, 26) = 1.09$, $p = 0.403$).

Participants recruited through AMT consisted of a total of 32 females and 33 males, all of whom were native speakers of US English. No participant chose to withhold information when given the option (for gender, age, education, and occupation). As expected, participants recruited through AMT were not just students, with 40 % being employed full-time, 17 % employed part-time, 9 % being homemakers, 3 % having retired, 17 % studying full-time, and 14 % others. Likewise, the spread of age ranges reflects a non-student population, with 5 % being under 20, 51 % aged 20 to 29, 26 % aged 30 to 39, 9 % aged 40 to 49, 8 % aged 50 to 59, and 2 % aged 60 to 69. As for education, 2 % left full-time education before high school, 12 % completed high school, 40 % had some college education, 35 % had a bachelor's degree, 9 % a master's, and 2 % a doctorate. Age groups and genders are distributed evenly across groups (age: Fisher's Exact test, $p = 0.616$; gender: Fisher's Exact test, $p = 0.929$).

None of the participants wore hearing aids and none self-reported problems with hearing. Only 8 % of AMT participants scored 10 or higher on the HHIA-S,

Table 5.2: *Significance of individual-level predictors (analysis of variance (ANOVA) model comparison, χ^2 test) for the GLMM of the low reverberation experiment*

	Lab		Amt	
Predictor	**AIC**	***p***	**AIC**	***p***
Baseline	1979	—	6066	—
System	2009	0.001	6148	0.001
Reverberation	1972	0.545	6168	0.001
System x Reverberation	1974	0.535	6074	0.002

but the WERs of two of them were the highest for their respective groups. However, a linear model predicting the transformed mean WER did not show this to have had a significant effect. Nor did the same model show any significant effect of synthetic speech listening frequency, age group, gender, whether participants described themselves as computer scientists, or their degree of left- or right-handedness ($R^2 = 0.13$, $F(8, 56) = 1.03$, $p = 0.425$).

5.3.2 Results

The results from the lab participants can be seen in the top of Figure 5.1, which shows traditional box plots (where the midline is the median value) of the mean WERs aggregated by system and level of reverberation with the mean value of each plot indicated by the × symbol. We had hypothesized that natural speech would fare better in reverberation than speech from either of the synthesis methods and Figure 5.1 would appear to support this hypothesis. Indeed, it *appears* to go further and show that NS and F-USEL become less intelligible with increasing reverberation, but that HTS enjoys an interesting *increase* in intelligibility in the reverberation at 250 cm, in much the same way (but in the opposite direction) that it did in additive noise. However, the range of WERs covered by the graph is very small and, unsurprisingly, a generalized mixed effects model built to predict the number of errors from system and its interaction with level of reverberation—represented by the formula given in Equation (5.5)—found no significant effects other than the difference between systems themselves (see Table 5.2)

$$\text{Error} \sim \text{System} * \text{ReverbLevel} + (1 + \text{System} \mid \text{Sentence}) + (1 \mid \text{Participant}) \quad (5.5)$$

The results from AMT participants (shown in the bottom of Figure 5.1) mirror those from the lab, albeit with more variance, as one would expect. As with the results from lab participants, the WER for HTS does not increase at the same rate as

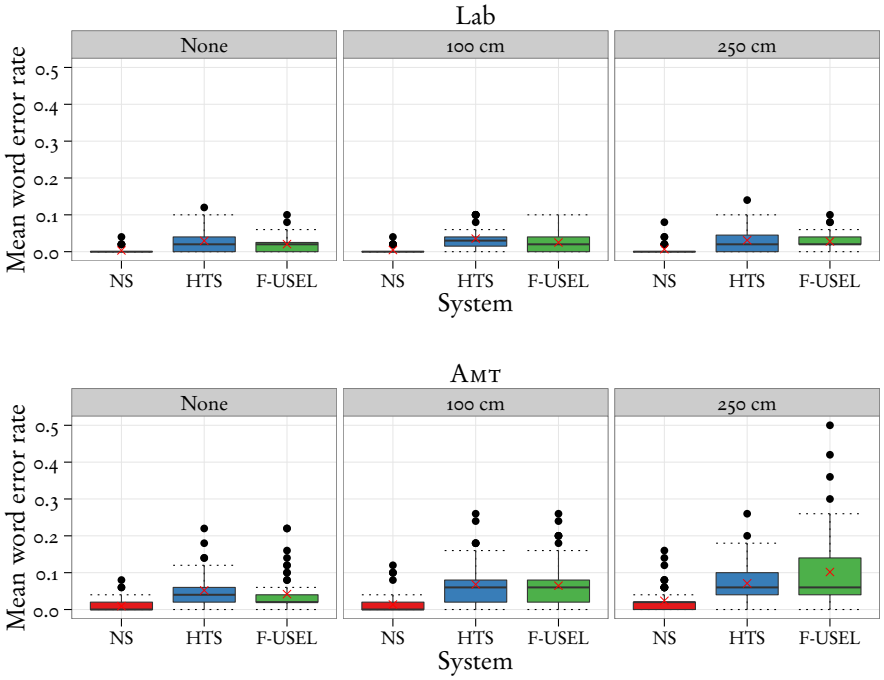

Figure 5.1: *Mean WER by system and level of reverberation with the lab and AMT*

for natural and unit-selection speech, although it stays more or less static rather than decreases. In the case of the AMT results, the level of reverberation and its interaction with system does become statistically significant, as Table 5.2 shows.

5.3.3 Discussion

Our first hypothesis was that an increased level of reverberation would result in a corresponding increase in the WER for both natural and synthetic speech. The initial results from the lab seem to contradict this and show no significant difference in WER at the levels of reverberation we used. However, the same stimuli heard by AMT participants did reveal a significant effect of both reverberation level *and* the interaction of reverberation level with speech system. The difference in these results perhaps confirms our conjecture that experiments conducted only in the listening lab do not provide a full picture of the likely performance of systems in the real world. Although the differences between HTS and F-USEL are not great, the results from AMT suggest that the lower variance encountered with HTS may make it a better choice for reverberant environments, particularly at distances greater than the critical distance, as in our 250 cm level. Moreover, even a small increase in intelligibility may be perceived as a major benefit to someone who has to use the system over an extended period. In any event, the results give a clear indication of the reduction in intelligibility to be expected when using synthetic rather than natural speech at moderate levels of reverberation. We expected the effect to be more marked in a

higher level of reverberation, such as in a railway station or church and investigated this in our next experiment.

We had wanted to compare AMT with the laboratory for investigating reverberation because we had hypothesized that any background noise present in the AMT listeners' environments could be a confounding factor in the results. However, Figures 5.1 shows that the relative rankings of the three systems tested are maintained when using AMT. Nonetheless, we wanted to establish the effect of any background noise in the AMT listeners' environments. For this reason, we had asked participants to report how noisy the environment was in which they had taken the listening test, along with a number of questions designed to elicit what other noise they could hear and to cross-check their answers. The bottom of Figure 5.2 replicates the bottom of Figure 5.1, but with results only from those participants we can be confident were in a quiet environment all of the time (41 of the 65 total AMT participants). We determined these by dividing participants into two groups: those who reported no noise whatsoever (so, as near as possible to lab conditions) and those who reported any form of extraneous noise. Although lab students should not have been able to hear any external noise, a number reported their environment as not being quiet all of the time. We suspect that they were reporting the reverberation in the stimuli, but to be sure we carried out the comparison only with those who had reported their environment being quiet all of the time (the top of Figure 5.2). It seems that any background noise the participants heard did not interfere significantly with their ability to comprehend the stimuli.

However, to verify this, we built a generalized linear mixed model (GLMM) to assess the extent of the effect of background noise on the results from AMT. The model was built using the binary division of participants into those in quiet and those in some form of noise described above. We then modelled the number of errors predicted by system, reverberation level, the binary value of quiet or non-quiet environment, and all possible interactions. An ANOVA comparison of the full model with a reduced model (with quiet and all its interactions removed) did not find the effect of noise on the results to be significant (χ^2 test $p = 0.02$). Even so, although we had already shown noise to have a significant effect on intelligibility in previous experiments, it is somewhat surprising that it seems as though the noise AMT participants heard was sufficiently loud as to affect intelligibility almost to the point of statistical significance even though they reported wearing headphones. The ability of noise to act this way may be related to the contrary effects of noise and reverberation. Whilst some reverberation seems to improve the intelligibility of HTS, noise has the opposite effect. It would seem prudent, then, when running reverberation experiments on AMT to enquire about the listening environment of

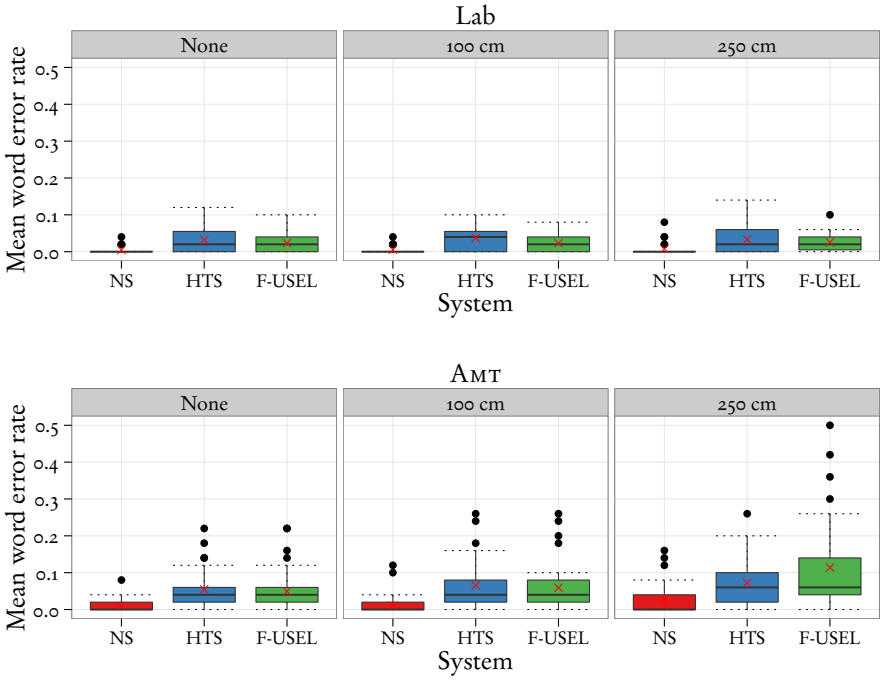

Figure 5.2: *Mean WER by system and level of reverberation with those lab and AMT participants who reported their environment quiet all of the time*

participants and to exclude any results where any extraneous noise was reported. The contrary effects of noise and reverberation had interesting implications for situations in which they occurred together. We report on our investigation of this in Chapter 6, but first we looked at the effects of a much greater level of reverberation.

5.4 High-level reverberation experiment

Having observed the slight improvement in intelligibility with HTS in low reverberation, we wanted to see what would happen with a much larger reverberation. Many of the prospective users of speech technology for reminders will wish to do so in an environment of high reverberation, such as a train station, church, or simply the stairwell of their building.

For this experiment, we introduced a binaural impulse response recorded for the Aachen impulse response (AIR) database [Jeub et al., 2009]. The impulse response we chose was recorded in the Aula Carolina, Aachen, a former church with a floor area of 570 m² and a high ceiling typical of those found in churches. We ran the experiment on AMT, as we had done for the previous experiment, using exactly the same materials and method, except that the Aula Carolina impulse response was used instead of that of the lounge at 100 cm used previously. This meant that participants heard stimuli with reverberation levels equivalent to: none; at 250 cm in the lounge of the Victorian

house; and that from the Aula Carolina.

We hypothesized that WERs would increase significantly for all speech forms in the increased reverberation, but that synthetic speech would suffer more than natural speech. We also hypothesized that HTS would not benefit significantly from the higher reverberation in the way that it had done at 250 cm in the lounge of the Victorian house.

5.4.1 Method

All participants—11 females and 23 males—were recruited via AMT, were native speakers of US English, and were reimbursed US$1 for their time. One participant chose to withhold information about age, gender, education, and occupation, whilst a further participant withheld information about occupation only.

The information provided showed that participants came from almost the full range of ages, education, and occupations. Ages included 6 % who were under 20, 40 % aged 20 to 29, 29 % aged 30 to 39, 11 % aged 40 to 49, 9 % aged 50 to 59, and 3 % aged 60 to 69. Education included 3 % who left full-time education before high school, 14 % who had completed high school, 37 % who had some college education, 29 % who had a bachelor's degree, 9 % a master's, and 6 % a doctorate. Occupations included 37 % who were employed full-time, 20 % employed part-time, 11 % were homemakers, 9 % were studying full-time, and 17 % others. Age groups and genders are distributed evenly across groups (age: Fisher's Exact test, $p = 0.641$; gender: Fisher's Exact test, $p = 0.911$).

Only 6 per cent of AMT participants scored 10 or higher on the HHIA-S, but most actually had the lowest WER for their group and a linear model predicting transformed mean WERs did not show this to have had a significant effect. Nor did the same model show any significant effect of a reported problem with hearing, synthetic speech listening frequency, age group, gender, whether participants described themselves as computer scientists, or how left- or right-handed they were ($R^2 = 0.39$, $F(11, 23) = 1.33$, $p = 0.27$).

5.4.2 Results

Figure 5.3 clearly shows the effect of the much increased level of reverberation in the Aula Carolina (AC) condition. As we had hypothesized, HTS *seems* to benefit from the 250 cm level of reverberation, as it did in the previous experiment, but loses the advantage at the level found in the Aula Carolina. However, using the same formula used for the previous experiment (Equation (5.5) on page 106), the results of GLMM

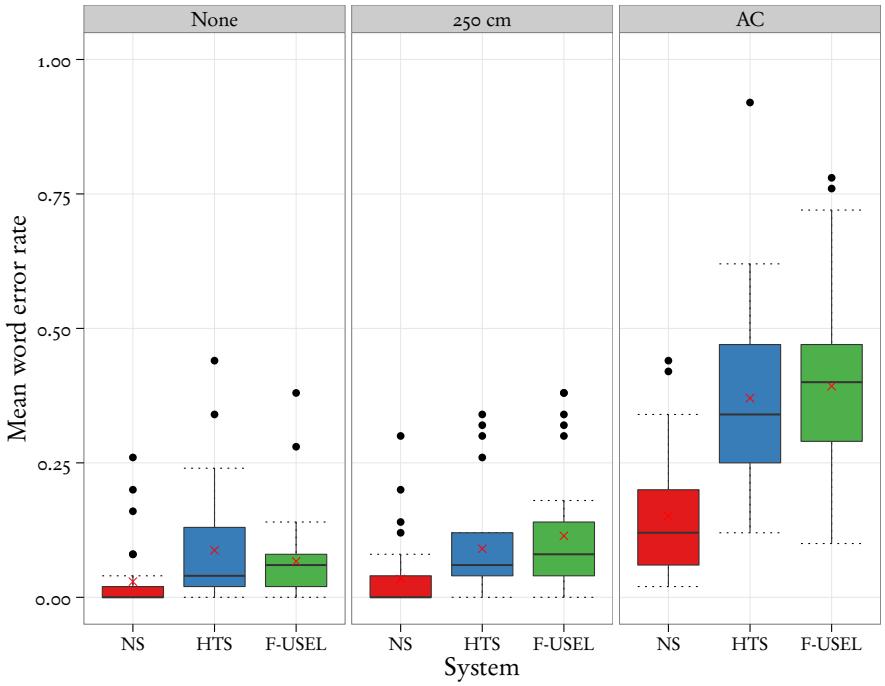

Figure 5.3: *Mean WER by system and level of reverberation*

Table 5.3: *Significance of individual-level predictors (ANOVA model comparison, χ^2 test) for the GLMM of the high reverberation experiment*

Predictor	AIC	*p*
Baseline	5886	—
System	6276	0.001
Reverberation	7156	0.001
System x Reverberation	5887	0.053

comparisons represented by Table 5.3 show that the interaction between system and reverberation was not significant in this experiment.

To assess the effect of noise on the results, we built a GLMM in the same way that we did for the previous experiment. There were 20 participants in lab-like conditions and 15 who reported some form of background noise. Again, we found the effect of noise not to be significant (χ^2 test $p = 0.32$).

5.4.3 Discussion

The levels of reverberation used in this experiment cover the range of reverberation levels likely to be encountered in day-to-day environments, from direct arrival, through well past the critical distance in a typical lounge, to a highly reverberant edifice. The WERs elicited are comparable to those from the speech-in-noise experiments in Chapter 4, ranging as they do from almost 100 % to 0 % intelligibility. We have,

therefore, achieved a sound baseline for future experiments whilst allowing for the higher WERs we would expect from combined noise and reverberation.

Since any noise the AMT participants experienced in their environment did not significantly affect the results, the levels of reverberation also seem well suited for experiments conducted using AMT.

5.5 Conclusion

In much the same way as additive noise, reverberation has received scant attention in speech synthesis research compared to research into the intelligibility of natural speech. Our desire to test synthesis systems in ecologically-valid forms and levels of reverberation prompted us to source existent binaural impulse response from academic projects. In order to assess their use, we developed a process for combining them with speech and carried out experiments in the lab and on AMT. We sought to establish the level of parity between lab and AMT results and the effect of low and high levels of reverberation on natural and synthetic speech.

Once again, our results showed that NS is more resilient to interference than either of the synthetic speech systems, confirming our belief that research in this area is still warranted.

Our belief in the utility of AMT is further supported by the results in Table 5.2 on page 106, which demonstrate that using AMT has allowed us to find a significant effect not found in the lab.

How and why reverberation increases WER is perhaps best exemplified by Figure 5.4. It shows how the speech sounds (particularly the first three formants) are carried forward in time and appear as 'smears' on spectrograms. The figure shows the same utterance, 'Kathy wants twelve red mugs.', with the four levels of reverberation used in the experiments for this chapter. Even with the reverberation at 100 cm in the lounge of the Victorian house, some smearing can be seen, which is particularly noticeable at about the 4 kHz level just before 2 s has elapsed. It also seems clear that the energy in the speech sounds is being dissipated, since the red areas are not as red in the low frequencies and the yellow in the high frequencies is not as prevalent. In fact, it seems that some of the high frequency sounds might be 'lost' completely, which would account for the increased difficulty in perceiving consonants in reverberation.

The difference between the spectrograms for 100 cm and 250 cm is not immediately obvious, although it is there, and probably accounts for why we found little difference between them in the first experiment. It does seem clear that including these two levels of reverberation in the same experiment is not particularly productive. On the other hand, the justification for including reverberation from the Aula Carolina

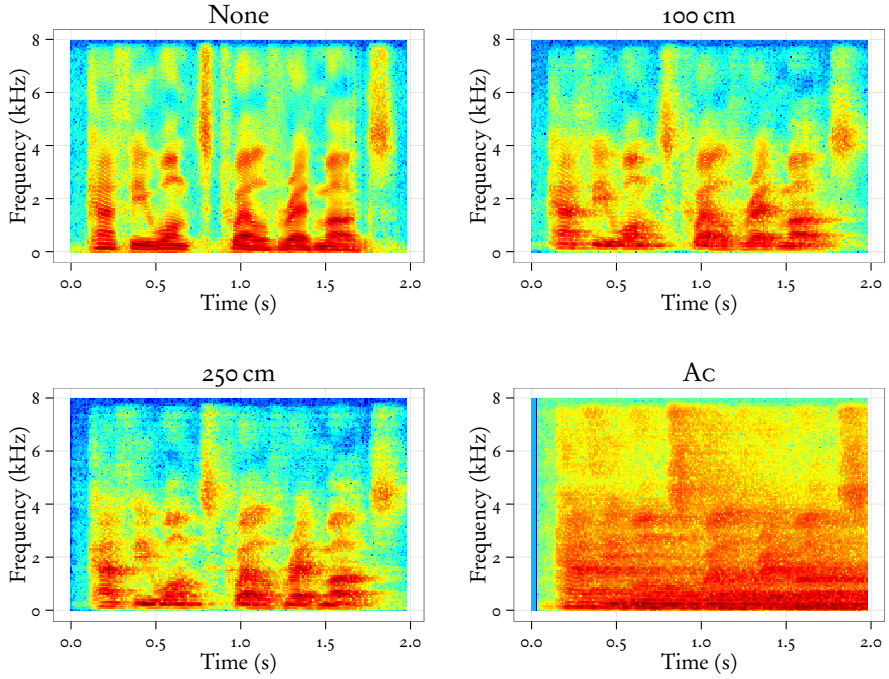

Figure 5.4: *Spectrograms showing effects of various levels of reverberation*

is obvious from the spectrogram alone. The much higher level of reverberation has fundamentally changed the soundscape and almost completely obliterated the periods of silence between the speech sounds. At first glance, it is difficult to see how any part of the utterance could be intelligible and the increased WERs experienced with this level of reverberation would seem to be expected. The use of the conditions, none, 250 cm, and AC would appear to be a good choice for experiments involving intelligibility in reverberation.

In experiments involving noise, we would expect the reverberation of the noise to exacerbate the interference with intelligibility, since the noise will also suffer from 'smearing' and would be expected to interfere with speech over a longer time frame than otherwise would be the case. We investigated the effect of noise and reverberation combined in our final experiment in Chapter 6.

CHAPTER 6

SYNTHETIC SPEECH IN NOISE AND REVERBERATION

ALTHOUGH it appears that there have not been any studies investigating the intelligibility of synthetic speech in noise and reverberation combined, some work has been carried out for natural speech, focused mainly on classrooms and hearing impairment. Since noise and reverberation individually affect the intelligibility of speech, their combined effect might be expected to do the same. In fact, both Nabelek and Mason [1981] and Payton et al. [1994] found that the combined effect was greater than the individual contributions.

In this chapter, we present the results of a large-scale experiment intended to examine the combined effect of noise and reverberation on modern speech synthesis and natural speech using the methodologies we developed for the purpose.

6.1 Method

In this experiment we wanted to assess the effect of simultaneous reverberation and noise on the intelligibility of natural and synthetic speech, using the systems, noises, noise levels, and reverberation levels used for previous experiments to allow for accurate comparisons between experiments. Our primary aim was to establish a baseline of intelligibility for modern systems in noise and reverberation and to confirm that experiments run using Amazon Mechanical Turk (AMT) would produce results similar to those in the lab.

We used the short, non-confusing, reminder-like Matrix sentences we had used for

previous experiments as we had shown that they provide the same relative rankings as semantically unpredictable sentences (SUSS) and would enable comparisons with previous experiments. The same three systems: natural speech (NS), the HMM-based Speech Synthesis System (HTS), and *Festival* unit-selection (F-USEL) that were used to create the stimuli for the experiments in Chapter 5 were used to generate the stimuli for this experiment.

Three reverberation levels were chosen: none; that found at 250 cm in the lounge of the Victorian house; and that from the Aula Carolina (AC) (all of which are described in Chapter 5). The 'none' level effectively gives us a control condition; the AC a maximal level; whilst 250 cm is the level at which we had observed an interesting interaction with HTS.

We elected to use noises that had previously been used in the ecologically-valid noises experiment described in Chapter 4, that is: none, chat, music, and both (the noises are described in detail in Section 4.3). As with no reverberation, no noise effectively gives a control condition and the other noises are those hypothesized (and indicated by AMT) to be most commonly found in domestic environments. The levels at which the noises would be presented were chosen such as to achieve signal-to-noise ratios (SNRS) of 5, −5, and −15 dB. These SNRS provide an evenly distributed difficulty level ranging from, approximately, 100 % to 50 % word error rate (WER) without reverberation.

The number of conditions for this experiment, therefore, would be 108, being 3 systems ×3 reverberation levels ×4 noises ×3 SNRS. Since we intended to have 10 sentences per condition, participants would have been required to listen to 1080 sentences, which would have taken them at least four hours. In order to make the experiment more manageable, we elected for a mixed design in which each participant would hear all three reverberation levels, all three SNRS, and *one* combination of system and noise. That is to say, reverberation and SNR effects would be measured within participants with system and noise being measured between participants. Thus, each participant would only need to listen to ninety sentences and a total of twelve participants would be required to ensure the coverage of each combination of conditions. However, in order to achieve a balance of sentences across conditions, three arrangements were generated per combination of system and noise, thus necessitating a minimum of thirty-six participants to ensure full coverage.

Each sentence, in each of the three systems, was normalized to the equivalent of 65 dB sound pressure level (SPL) and was then convolved with either a vector of a one and all zeroes for the condition without reverberation or the appropriate binaural impulse response (BIR) for the 250 cm and AC levels of reverberation. For each of the four noises (none, chat, music and both) a section equal in length to the speech was

selected at random. The selection was then adjusted to the same level as the speech and convolved appropriately for all three reverberation levels before being added to the speech at the required SNR. The SNR was achieved by adjusting the speech rather than the noise so that the correct level of speech would be presented in the conditions where there was no background noise.

Participants were recruited through AMT and the University of Edinburgh's recruitment service. Demographic data collected from all participants is summarized in Table 6.1.

Those participants recruited through AMT were all native speakers of US English and were reimbursed US$1 for their time. A total of 260 participants registered to complete the experiment with only 77 actually completing all sections. Whilst it is not possible to know why such a high number chose not to complete, comments from some suggest that the difficulty of listening to the stimuli relative to the remuneration may have been an issue. Just under 8 per cent of AMT participants scored 10 or higher on the Hearing Handicap Inventory for Adults Screening Version (HHIA-S). A linear model predicting appropriately transformed mean WERS did not show this, synthetic speech listening frequency, age group, gender, whether participants described themselves as computer scientists, or how left- or right-handed they were to have had a significant effect ($R^2 = 0.08$, $F(9, 67) = 0.68$, $p = 0.724$). One participant wore hearing aids in both ears and was removed from our analysis as we cannot be certain what facilities the hearing aids had for compensating for noise or reverberation. No participant self-reported a problem with hearing.

Participants recruited for the lab part of the experiment were paid £8 for their time. Just under 8 per cent of the lab participants scored 10 or higher on the HHIA-S and a further two reported minor hearing problems. None of their mean WERS gave cause for concern and a linear model predicting appropriately transformed mean WERS did not show any significant effect. Neither did the same model show any significant effect ($R^2 = 0.11$, $F(9, 61) = 0.86$, $p = 0.565$) of synthetic speech listening frequency, age group, gender, or whether participants described themselves as computer scientists.

6.2 Results

This experiment was built on those from Chapters 4 and 5 by including conditions with both noise and reverberation. It also included conditions with either noise or reverberation only so that we could make comparisons with previous experiments.

The overall results of the AMT and lab parts of the experiment are summarized in Figures 6.1 and 6.2 respectively. The pattern of results is as would be expected with increasing noise and reverberation levels resulting in higher WERS and NS performing

Table 6.1: *Demographic data collected from AMT and lab participants*

		AMT		Lab		TOTAL	
		n	%	n	%	n	%
Gender	female	44	57	37	51	81	54
	male	33	43	35	49	68	46
Age group	18–29	28	36	68	94	96	64
	30–49	33	43	2	3	35	23
	50+	16	21	1	1	17	11
	withheld	0	0	1	1	1	1
Education	bachelors	32	42	8	11	40	27
	doctorate	2	3	0	0	2	1
	high school	3	4	26	36	29	19
	masters	10	13	15	21	25	17
	some college	30	39	22	31	52	35
	withheld	0	0	1	1	1	1
Occupation	employed full-time	33	43	0	0	33	22
	employed part-time	14	18	2	3	16	11
	homemaker	8	10	0	0	8	5
	other	7	9	4	6	11	7
	retired	4	5	0	0	4	3
	student full-time	9	12	65	90	74	50
	withheld	2	3	1	1	3	2
Computer scientist	no	66	86	42	58	108	72
	yes	11	14	30	42	41	28
Work in speech technology	no	75	97	72	100	147	99
	yes	2	3	0	0	2	1
Listening Frequency	at least once a wk	12	16	14	19	26	17
	at least twice a yr	29	38	22	31	51	34
	not sure	6	8	7	10	13	9
	rarely or never	30	39	29	40	59	40
Headphones	earbuds	29	38	2	3	31	21
	full ear	22	29	69	96	91	61
	in ear	11	14	0	0	11	7
	on ear	15	19	1	1	16	11
Headphone features	noise cancelling	10	13	4	6	14	9
	none known	60	78	58	81	118	79
	sound isolating	3	4	4	6	7	5
HHIA-S total	10 or above	5	6	6	8	11	7
	below 10	72	94	66	92	138	93
Noisiness	noisy most time	1	1	5	7	6	4
	quiet all time	61	79	66	92	127	85
	quiet most time	15	19	0	0	15	10
	equal noise quiet	0	0	1	1	1	1
Browser	Chrome	37	48	0	0	37	25
	Firefox	17	22	0	0	17	11
	IE	9	12	2	3	11	7
	Mozilla	2	3	0	0	2	1
	other	1	1	0	0	1	1
	Safari	11	14	67	93	78	52
	Opera	0	0	3	4	3	2
Experience of stimuli	hard to type	1	1	0	0	1	1
	usually all words	7	9	10	14	17	11
	usually most words	30	39	43	60	73	49
	very hard	39	51	19	26	58	39
	TOTAL PARTICIPANTS	77	100	72	100	149	100

better than the two synthetic systems in every condition. Similarly, the results from AMT mirror very closely those from the lab, but with more variance, as we have come to expect. It seems that, once again, AMT has proved to be as good as the lab for investigating the relative performance of speech in noise and reverberation. The results of analysis of variance (ANOVA) comparisons of generalized linear mixed models (GLMMs) are shown in Table 6.2. The fully-specified model was built with the individual-level predictors System, Chat (whether present), Music (whether present), SNR, and Reverberation and the person and sentence identifiers as group-level predictors. The formula used is given in Equation (6.1). It can be seen that in both the lab and AMT parts of the experiment, all of the individual-level predictors were significant.

$$\text{Error} \sim \text{System} * \text{ReverbLevel} * \text{Chat} * \text{Music} * \text{SNR} + (1 \mid \text{Sentence}) + (1 \mid \text{Participant}) \tag{6.1}$$

Our results show the expected interaction between noise and reverberation: as SNR increases, the negative effect of reverberation on intelligibility is amplified. The three systems also differ in their susceptibility to the combined effects of noise and reverberation. At a SNR of 5 dB, HTS maintains its slightly better performance relative to F-USEL as reverberation increases, but at the SNR of −15 dB, the advantage disappears completely. The type of noise, on the other hand, does not appear to influence the effect of reverberation on intelligibility.

In comparison with previous experiments, the ecologically-valid noises experiment presented in Section 4.3 is most similar to the lab part of this experiment, without reverberation (that is, those columns labelled 'None' under the label for noise type in Figure 6.2). For ease of comparison with Figure 4.4 on page 89, Figure 6.3 shows only those results from the lab part of this experiment that do not include reverberation.

On comparing the two, it is clear that this experiment resulted in WERS that, across the board, are somewhat higher. In fact, they are more in line with what we might have expected with WERS exceeding 50 % at −15 dB as Rhebergen and Versfeld [2005], and our own results from the baseline experiment (see Figure 3.1 on page 46), suggested they would do. The implication is that, as we pointed out in Section 4.3.2, the WERS for our ecologically-valid noises experiment were unexpectedly low. In Section 4.3.2 we postulated that the reason for the lower WERS was the more fluctuating nature of the speech, which included pauses. However, these results contradict that suggestion and lead us to believe that the true reason is the fact that, in the previous experiment, the speech stimuli were presented from speakers in front of the listener and the noise from behind, which perhaps allowed for better discrimination of the speech from the

Figure 6.1: *Mean* WER *by system, background noise, level of reverberation, and* SNR *(dB) with* AMT *participants*

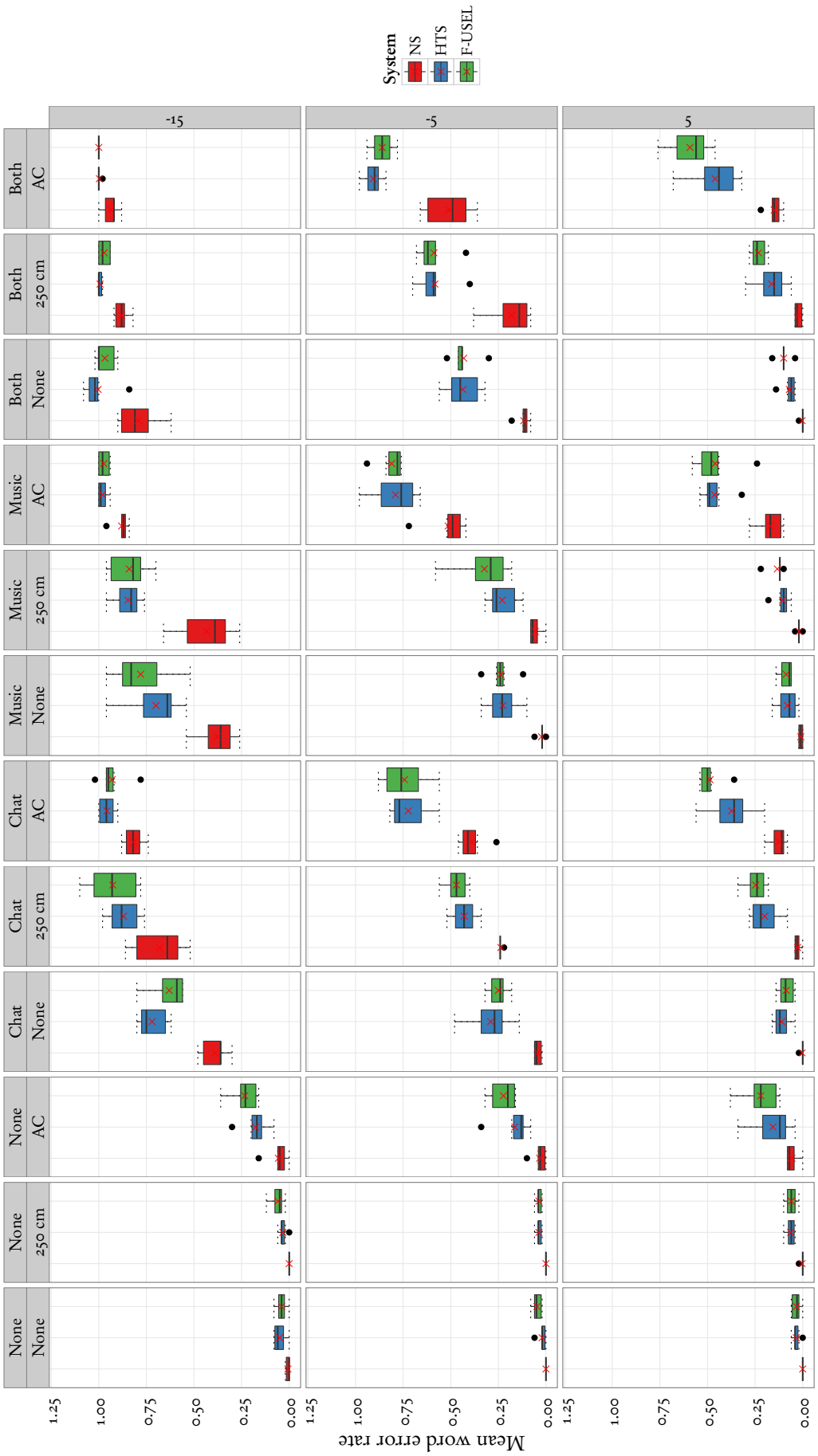

Figure 6.2: *Mean WER by system, background noise, level of reverberation, and SNR (dB) with lab participants*

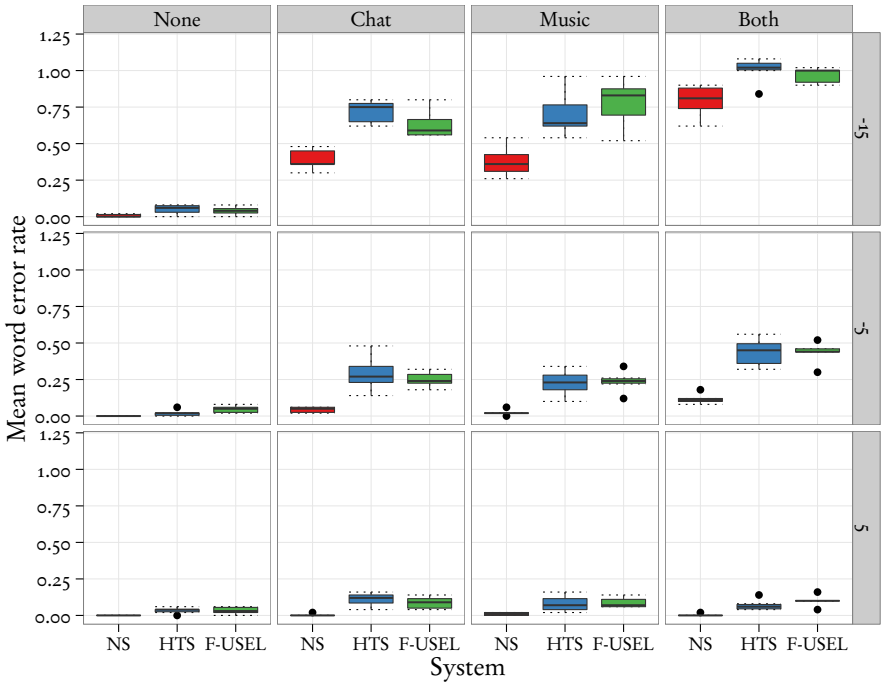

Figure 6.3: *Mean WER by system, background noise, and SNR (dB) of those lab results without reverberation*

noise. We had chosen that arrangement on the assumption that it would be the most likely scenario in the home environment, as it seemed logical that more often than not a sound source, such as a radio or television (TV) would be placed in front of the listener and any noise would, consequently, be more likely to be behind, or at least to the side of, the listener.

The high-level reverberation experiment in Chapter 5 was conducted solely with AMT participants and is, therefore, most similar to the AMT parts of this experiment without any background noise. However, unlike the previous experiment, this one includes three SNRs, none of which provides stimuli identical to those used previously. In order to make a reasonable comparison, we have chosen to use only those stimuli at a SNR of −5 dB, being the middle of the three and as near as possible to that used in the previous experiment. This subset of results is presented in Figure 6.4, which is clearly very similar to those in Figure 5.3 on page 111.

6.3 Discussion

The spread of results shown in Figures 6.1 and 6.2 suggest that our choice of materials for this experiment allows for testing across the full range of listening conditions with appropriate WERs being returned ranging from 0 to greater than 1. Even where we might be concerned about floor and ceiling effects creeping in, the relative perfor-

Table 6.2: *Significance of individual-level predictors (ANOVA model comparison, χ^2 test) for the GLMM of the noise and reverberation experiment*

	Amt		Lab	
Predictor	**AIC**	***p***	**AIC**	***p***
Baseline	20568	—	15995	—
System	21072	0.001	16513	0.001
Reverberation	22387	0.001	18031	0.001
Chat	20987	0.001	16520	0.001
Music	20875	0.001	16433	0.001
SNR	23884	0.001	20423	0.001
System x Reverberation	20568	0.001	15995	1.000
System x Chat	20568	1.000	15995	0.001
Reverberation x Chat	20568	1.000	15995	1.000
System x Reverberation x Chat	20568	1.000	15995	0.001
System x Music	20568	1.000	15995	1.000
Reverberation x Music	20568	0.001	15995	1.000
System x Reverberation x Music	20568	1.000	15995	0.001
Chat x Music	20568	1.000	15995	1.000
System x Chat x Music	20568	0.001	15995	1.000
Reverberation x Chat x Music	20568	1.000	15995	1.000
System x Reverberation x Chat x Music	20575	0.005	15994	0.099
System x SNR	20568	1.000	15995	1.000
Reverberation x SNR	20568	0.001	15995	0.001
System x Reverberation x SNR	20568	0.001	15995	0.001
Chat x SNR	20568	0.001	15995	1.000
System x Chat x SNR	20568	1.000	15995	1.000
Reverberation x Chat x SNR	20568	1.000	15995	0.001
System x Reverberation x Chat x SNR	20568	1.000	15995	0.001
Music x SNR	20568	1.000	15995	1.000
System x Music x SNR	20568	1.000	15995	0.001
Reverberation x Music x SNR	20568	1.000	15995	1.000
System x Reverberation x Music x SNR	20568	1.000	15995	1.000
Chat x Music x SNR	20568	1.000	15995	1.000
System x Chat x Music x SNR	20568	0.001	15995	1.000
Reverberation x Chat x Music x SNR	20568	1.000	15995	1.000
System x Reverberation x Chat x Music x SNR	20564	0.389	15993	0.201

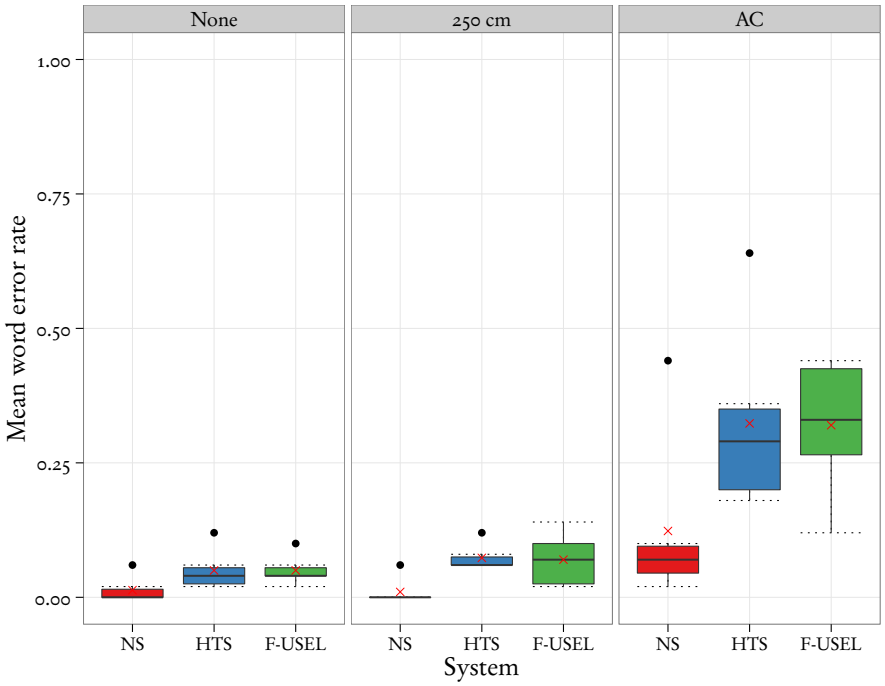

Figure 6.4: *Mean WER by system and level of reverberation of those AMT results without noise and with speech at the equivalent of –5 dB SNR*

mance of systems can still be determined.

Figure 6.2 allows us directly to compare the effect on intelligibility of noise versus reverberation. The top two rows of labels indicate the noise and reverberation levels, respectively. The three leftmost columns of the graph show the effect of increasing levels of reverberation on speech without any background noise. Taking the −15 and −5 dB SNR conditions (that is, where the level of the noise exceeds that of the signal) in these three columns and comparing them with the columns with a noise but no reverberation ('None' in the second row of labels), it is evident that noise alone has a greater impact on intelligibility than even the highest level of reverberation. However, when the position is reversed and the SNR is positive, the situation is less clear-cut and the highest level of reverberation (AC) does exceed the disruption caused by noise.

We can also see that, as Nabelek and Mason [1981] and Payton et al. [1994] found with natural speech, reverberation and noise together affect intelligibility more than either of them alone. Looking at the conditions with noise in Figure 6.1 (those columns headed, 'Chat', 'Music', or 'Both'), where the SNR is 5 and −5 dB, the WERS are greater when noise and reverberation are combined than they are with either noise or reverberation alone. When the SNR is −15 dB, the position is not as clear cut as the WERS are near ceiling levels, nonetheless they do increase slightly. The reason for this is probably that when a fluctuating noise is subject to reverberation, the smearing of the noise reduces the release from masking—that is, the gaps from which

listeners can glimpse the speech stimulus—and it effectively becomes static in nature. This is apparent from the spectrograms shown in figure 6.5. Each row represents a separate utterance at each of the three SNRs (5, −5, and −15 dB) and the columns represent increasing reverberation, from none to that from the AC.

For conditions without reverberation, the WERs in this final experiment were higher than those for the similar experiment in Chapter 4 when they might have been expected to be statistically the same, or even perhaps slightly lower, given the findings of Raitio et al. [2012]. The fact that they were closer to our expectations than the previous experiment highlights the anomalous nature of that experiment. It seems that presenting speech and noise separately in front and rear speakers does not elicit the same response as a mixed mono or stereo signal presented through headphones. (Or, probably, a mixed signal presented through a multichannel speaker set-up as used by Raitio et al. [2012].) The discrepancy highlights the effect experimental set-up can have on results and, consequently, how important it is to match it with conditions in the real world.

For conditions without noise, the results were as we would have expected, so that, overall, we can be confident that the results from this experiment concur with previous experiments and thus provide a true picture of the effects of the combination of noise and reverberation relative to those experiments.

It seems that in the most difficult listening conditions participants achieved WERs greater than 100 %. The reason appears to be that our chat background noise included discernible words and some participants reported words from the noise rather than the signal. This is unfortunate, since it undermines the advantage that Matrix sentences have of having a fixed format and vocabulary set, such that they tend not to suffer as much from the spurious insertion of extra words that were not actually presented. Indeed, this was the first experiment we had carried out that returned WERs in excess of 100 %. Clearly, this would not happen in experiments where the background noise included speech-shaped noise, but no actual words, and we intend to investigate ways of achieving this, without compromising the realistic nature of the background noise, in future work.

6.4 Conclusion

The final experiment presented in this chapter brought together all our experimental methodology to establish a baseline for performance of NS, F-USEL, and HTS in noise and reverberation. The applicability of the methodologies to experiments run on AMT was confirmed through a close correlation with results from the lab.

To the best of our knowledge, this is the first systematic investigation of the

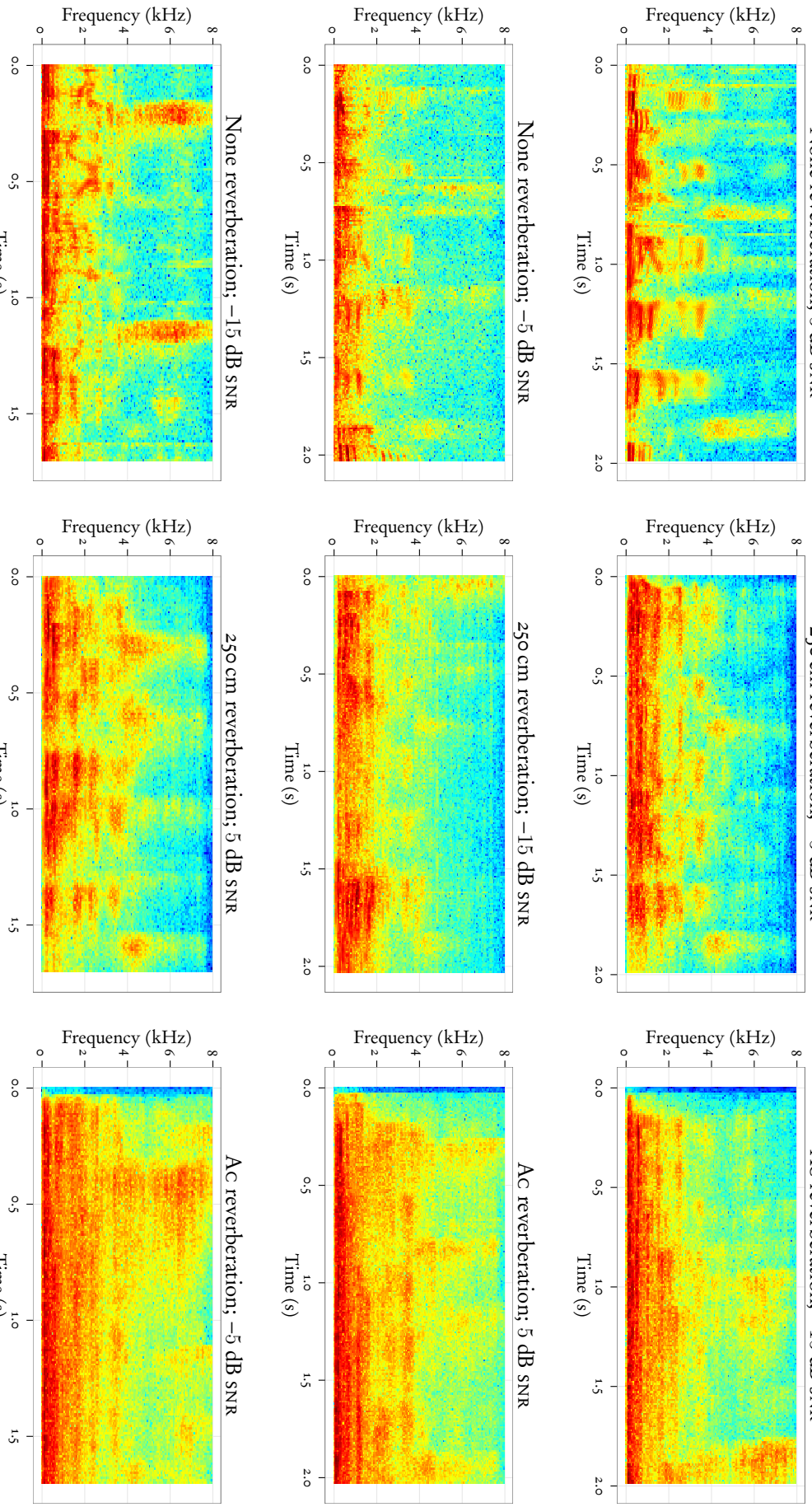

Figure 6.5: *Spectrograms showing the effect of noise and reverberation at the nine combinations of SNR and reverberation level*

combined effect of noise and reverberation on the intelligibility of synthetic speech. Our results mirror existing findings, for natural speech, that there is an interaction between noise and reverberation. As we have seen earlier, under bad listening conditions (high reverberation, low SNR), synthetic speech intelligibility degrades far more than the intelligibility of natural speech. We also found that the effect of reverberation is higher on HTS than on F-USEL at the lowest SNR.

CHAPTER 7

CONCLUSION AND FUTURE WORK

WE HAVE developed a rigorous methodology for the evaluation of speech synthesis and applied it to provide a better understanding of the effects of noise—both additive (in the form of background noise) and multiplicative (in the form of reverberation)—on the intelligibility of synthesized speech that can be applied immediately to state-of-the-art synthesizers, such as the HMM-based Speech Synthesis System (HTS) and *Festival* unit-selection (F-USEL) and, more generally, to the development of future systems.

In Chapter 2, we presented a review of the literature covering studies of the evaluation of synthetic speech. That review revealed a lack of recent research into the intelligibility of modern synthesis systems in noise, reverberation, or a combination of the two. Moreover, those studies that did exist had been carried out in a laboratory with undergraduate students and tended to rely on noises and stimuli that do not occur in the real world. As a result, we sought to develop an experimental methodology that could bring to bear recent developments onto the assessment of synthetic speech intelligibility.

In Chapter 3, we set out the components of the methodology: Amazon Mechanical Turk (AMT), Matrix sentences, and ecologically-valid noises, each of which was then examined and evaluated for use in intelligibility, firstly with synthetic and, ultimately, with natural speech.

Having satisfied ourselves that the new techniques elicit comparable results, but with their own advantages, we went on to confirm their efficacy and to test the intelligibility of two state-of-the-art synthesis systems with their natural speech equivalent in noise, reverberation, and the two combined in Chapters 4, 5, and 6.

7.1 Contributions

The primary contribution of this work is an in-depth investigation of the intelligibility of modern text-to-speech synthesis systems in settings that are more ecologically valid than standard laboratory approaches. For this purpose, we introduced two methodological innovations: the use of AMT to move beyond the laboratory environment; and the use of Matrix sentences as a more accessible alternative to semantically unpredictable sentences (SUSS), which also allows us to bridge intelligibility tests and audiological assessments.

In order to increase ecological validity, we used a standardized, speech-shaped, fluctuating noise, real chat, and music as our background noises. We also systematically investigated the effect of reverberation, which has been neglected in the literature on synthetic speech evaluation so far, even though reverberation is present in all real-life listening situations that do not involve headphones.

For data analysis, we used generalized linear mixed models (GLMMS), which are emerging as the statistical analysis method of choice in psycholinguistics and speech perception.

7.1.1 Amazon Mechanical Turk

The use of a crowdsourcing platform in the work for this thesis was novel and resulted in the first published presentation of its use for the evaluation of synthetic speech intelligibility [Wolters et al., 2010]. This aspect of the thesis has probably had the greatest impact in that the use of crowdsourcing, in the form of AMT, has subsequently been adopted by the organizers of the *Blizzard Challenge*, synthetic speech researchers (for example, Watts et al. [2010] and Black et al. [2012]), and commercial providers of speech synthesis systems. It is now routinely used for the comparative assessment of synthesis systems, since it is cheaper and quicker than running laboratory experiments and, as we have shown, allows for the recruitment of a larger and broader-based cohort of participants enabling the detection of even small, but significant, effects.

7.1.2 Matrix sentences

Similarly, prior to our use, we are not aware of any synthetic speech intelligibility studies employing Matrix sentences. They have since been used more generally for the assessment of objective measures of intelligibility [Valentini Botinhão et al., 2011a]. Matrix sentences have the advantage that they are multidisciplinary in that they confer a link with audiological assessment, for which they were developed. More

broadly, they enrich synthetic speech research by adding a set of stimuli that can be generated in large number and that have the benefit of being realistic and non-confusing for situations where this is a particular concern, such as with older or cognitively-impaired participants.

7.1.3 Ecologically-valid noises

Our results clearly show that using different types of ecologically-valid noises and settings affect the intelligibility of synthetic speech, and that different approaches to speech synthesis may vary in their robustness to degradation.

Since the aim of the thesis was to evaluate speech synthesis intelligibility, we did not perform an exhaustive evaluation of different types of noises. Instead, we initially took our lead from audiology and used a standard speech-shaped fluctuating noise. Our choice of more realistic background noises was motivated by a post-hoc analysis of AMT data, which revealed significant effects of chat and music. In order to establish a stable baseline for other researchers, we selected chat and music samples that did not vary much in intensity. In order to assess the effects of reverberation, we chose realistic impulse responses that came from the home and from a standard, highly reverberant environment.

7.2 Impact

The investigation into AMT detailed in Section 3.6 was published as a workshop paper [Wolters et al., 2010]; data collected from the experiments carried out for Chapter 3, and other assistance, contributed to a conference paper [Wolters et al., 2011]; and work carried out during the PhD project, but not presented in this thesis, contributed to a third paper [Wolters et al., 2012].

The most obvious beneficiaries of this work are those working in synthetic speech research, since not only will they be provided with an understanding of how up-to-date systems perform in noise and/or reverberation, but, perhaps more importantly, they will be left with a collection of methods that can be used in the assessment of future systems as they evolve. However, the results of the research will be of wider interest than just to those involved in speech synthesis.

Since the research was undertaken in the context of a project funded by the Engineering and Physical Sciences Research Council (EPSRC) investigating reminders in the home environment, the work benefits from using conditions that might be found in everyday environments whilst maintaining the rigour that comes from using an experimental methodology. Therefore, researchers and practitioners in the field

of human-computer interaction (HCI) and, ultimately, anyone who uses synthetic speech in noisy or reverberant conditions will have the information required to make informed decisions about the interfaces they develop.

In the meantime, insight into how background noise and reverberation affect modern (and future) synthetic speech might allow the choice of an appropriate system, or modifications that might be made to a system, in order to improve intelligibility.

Clearly, the impact set out above could be deemed to have been achieved simply by providing an analysis of synthetic speech in noise and reverberation and the validated methods required to support that analysis. A more substantive measure of success would be that the outcomes were taken up and used by the wider research community. In the event, many of them have been. Many language researchers are now using AMT, including the annual *Blizzard Challenge*, following, as a prelude to this thesis, the publication of the paper describing its use [Wolters et al., 2010].

Following its successful use in experiments for this thesis, the International Collegium of Rehabilitative Audiology (ICRA) noise was incorporated into a test of synthetic speech in noise for the *Blizzard Challenge* 2010 [King and Karaiskos, 2010] and subsequent research papers [Cooke et al., 2013; Valentini Botinhão et al., 2013].

Similarly, after establishing that Matrix sentences could, indeed, be used in place of SUSs and elicited the same *relative* word error rate (WER) scores, they were taken up by others, for example, Valentini Botinhão et al. [2011b,a]

7.3 Future work

As with any work of this nature, there remains much to be done, either because it was outside the scope of a single PhD or because of limitations on time or resources. Some areas for future work are laid out below.

7.3.1 Amazon Mechanical Turk

One of the advantages we have seen with using AMT is the broadening of the range of participants being recruited, so that we are no longer restricted primarily to young, full-time undergraduate students with the bias that inevitably implies. However, although AMT participants are more representative of the wider population than lab recruits, they are still not fully representative of the population [Ipeirotis, 2010]. This leaves the possibility that AMT has its own bias because of the demographics of the subset of the population who use it. Further work on the exact demographics of its users—from the basic factors, such as age and occupation, to more esoteric factors, such as their motivation for choosing to take part in experiments is needed. We might

find, for example, that blue-collar workers suffer more from hearing loss because a high proportion of them work in noisy environments.

As a result of the publication of our work on using AMT for experimental studies, the data collection for the *Blizzard Challenge* of 2010 and 2011 [King and Karaiskos, 2010, 2011] contained a sizeable number of judgements collected via AMT, which, to our knowledge, have yet to be fully analysed and published. The relatively large number of participants recruited for the challenges and the direct comparability of the AMT and non-AMT parts would make their analysis a useful addition to our own work and a useful comparison with it.

In August 2014, AMT tightened its restrictions on its use by non-US companies and residents, which could limit its accessibility to researchers in other parts of the world without a US-based partner. There are crowdsourcing services other than AMT, but their equivalence to the lab would need to be ascertained before they could be used for large-scale studies.

A related area of potential work would be the analysis of previous years' data from the *Blizzard Challenge* using GLMMs in order to make comparisons with the statistical analysis currently used.

7.3.2 Matrix sentences

Additional work could consider why Matrix and semantically unpredictable sentences perform differently—particularly why a crossover does, or does not, occur in the performance of HTS and F-USEL. Is it simply that Matrix sentences make semantic sense, or is it something more prosaic, for example that they differ in the distribution of consonants in their make-up?

A more pragmatic concern for synthetic speech researchers is whether the use of SUSs is tuning synthesis systems for them rather than improving the synthesis of real-world utterances. We have mentioned this possibility several times and possibly seen some evidence for it (such as the straight, parallel lines for both F-USEL and HTS in Figure 4.7 on page 95), but have not addressed it directly. We could develop hypotheses to explore whether SUSs are tuning systems or whether, in fact, the parallel lines which SUSs generated suggest they make a level playing field and are actually better for assessing intelligibility. An in-depth evaluation may also assist in answering the question, raised in Section 4.4.3, of whether the crossover in the performance of F-USEL and HTS is a product of whether the stimuli have semantic meaning.

In Section 3.7 we investigated and discussed the learning effects associated with Matrix and semantically unpredictable sentences and showed how the effect was larger for Matrix sentences. What we did not explore in any depth was why this might be.

There could be several reasons: the restricted vocabulary size, which might result in participants learning to recognize possible words; the high-frequency vocabulary selected, which makes word recognition easier; and the simple, highly predictable sentence structure. Further psycholinguistic research is required to establish to what extent each of these factors contributes to the learnability of Matrix sentences and how they affect the quality and validity of intelligibility judgements.

Part of our motivation for using Matrix sentences was to bridge the gap between the synthetic-speech- and audiology-research communities. We had hoped to build on this by including a speech reception threshold (SRT) test as a measure of intelligibility into our studies, since this is more common in audiology than the WER measure universally used in synthetic speech research. Unfortunately, it proved not to be possible to add our background noises to the standard tool for carrying out the requisite up-down procedure.

Nonetheless, there remains a body of research in the speech synthesis community orthogonal to that of the audiology community, but tantalizingly similar. For example, George et al. [2008] carried out an experiment to establish SRTs of natural speech in combined noise and reverberation with a number of similarities to our own. A rough-and-ready interpolation of our WERS reveals intriguingly similar SRTs, but there is no procedure available to make a direct comparison between the two experiments (or with others that used the SRT as a measure). A validated and robust procedure to convert WER scores to SRT signal-to-noise ratios (SNRS) would, in itself, be a productive area of research and would have been particularly beneficial to us in grounding our last experiment in the wider academic literature.

7.3.3 Ecologically-valid noises

The ecologically-valid noises we used only represent two types of noise: chat and music. The chat was chosen to be a relatively boring conversation and the music to be a relatively calm piece of Baroque. One obvious extension to our paradigm would be to devise a rigorous sample of different kinds of noises and reverberation levels that reflect typical environments in which synthetic speech is played. Assessing the effect of these different noises on intelligibility is very difficult in a traditional laboratory setting, but much easier when crowdsourcing. In a crowdsourcing design, noise type can be treated as a between-participant variable, and it is comparatively easy to recruit sufficient participants per noise.

An improvement that could be made to the way that the noises are used—rather than the noises themselves—is how they are combined to achieve the required SNR. We took great care to ensure that our noises were as natural and as representative

of real life as possible and were meticulous when combining them to ensure parity between stimuli. In the future, if a synthesis system were developed that could alter its output in the way that a human does when noise is present, it would be necessary to add the noise into the auditory scene, as was done for the Computational Hearing in Multisource Environments (CHiME) project Christensen et al. [2010], which would necessitate much more in terms of time and resources than was available for the PhD.

As we saw in Chapter 6, the use of ecologically-valid noises may also require a rethinking of assessment measures. In the most difficult listening conditions, where the background noise included speech, participants were including words from the background noise in their transcript of the speech stimuli. Traditional WER measures fail to capture this interference. A new measure is required that assesses both the degree to which the original signal was perceived (reflected in a corrected WER) and the degree to which the background noise was confused with the foreground speech.

7.3.4 Methodological improvements

In our baseline experiment, we equalized the durations of the HTS stimuli with those of F-USEL in order to obviate any bias caused by the length of the stimulus in the different systems. This practice continued in the following experiments for synthetic speech, but not for natural speech (NS). The consequence is that it is possible that the longer durations of the NS stimuli will have had a bearing on the WERs they elicited, particularly in conditions that included reverberation, since the longer the stimulus, the less reverberation there will be at the end of it. It would be prudent to investigate whether this was the case and, if so, ways of achieving equality of durations between all systems without adversely affecting the quality of any of the stimuli. Although HTS lends itself well to having the duration of an utterance varied, relying as it does on statistical parameters for generation rather than recordings of speech, we should also investigate whether doing so had any adverse effect.

When running our experiments, we provided a small number of practice stimuli so that participants could become familiar with the voice being used. In many real-life scenarios, such as reminder systems or robot companions, familiarization will occur over a much longer time frame. Where such a use were envisaged, it would be advantageous to carry out a longitudinal study to compare synthesis systems over a much longer period.

7.3.5 Different listener groups

The research presented here should be seen as a starting point from which further extensions could be made. For example, even though we broadened the range of

participants who took part, by using AMT, we have not done any work with older people or those with hearing or cognitive impairment. These groups are often the target audience for speech synthesis and have their own particular needs. Nor have we included any work on the effect of modifications to the speech stimuli, such as using clear or Lombard speech.

Moreover, synthetic speech has always had benefits for a range of applications, especially in situations where other modes of interaction are not available or require augmentation; either because of environmental conditions, or because of a disability or personal preference on the part of the user. Indeed, for some situations, including the provision of reminders in the home environment, synthetic speech has been shown to be one of the most effective means of delivery [McGee-Lennon et al., 2007].

The importance of reminder systems and alarms in hospitals, transport, process control, and industrial applications is well accepted [Noyes et al., 2006]. They can, though, be just as important for safety in the home environment, for example, alarms warning of appliances left on or doors left unlocked. Additionally, health care can be enhanced; for example, reminder charts for patients being discharged from hospital can significantly improve compliance with their drug regimen [Raynor et al., 1993]. Even more mundane reminders for appointments or events, mealtimes, toileting, or even just the time of day can greatly enhance the quality of life for someone with a cognitive impairment [Haigh et al., 2006]. The provision of reminders has the potential to help not only the increasing number of older people in society, but also people of any age who have cognitive or sensory impairments and, for this reason, has received some attention in the literature.

The use of technologies such as mobile devices and wireless networking raise exciting possibilities in presenting reminders. Presently, there is a plethora of systems being developed and trialled to assist people in the home environment, many of which include some form of reminder element. These systems are extremely heterogeneous varying from large scale intelligent homes [Perry et al., 2004; Nugent et al., 2008], through health-care systems [Zhou et al., 2010], to using pagers to provide self-reminders [Aldrich, 1998].

Surprisingly, although speech is widely used in products for people with visual impairment [RNIB, 2010], it has received little attention from the research community on its efficacy in presenting reminders. Nonetheless, the studies that have specifically focused on spoken reminders have shown that they can be beneficial. In addition to the obvious fact that a visual display is not required, the benefits have been summarized [Lines and Hone, 2002, 2003a] as follows:

- the majority of older adults are able to understand speech

- literacy is not required
- those with visual impairment can access information that would be difficult or impossible to access otherwise
- information can be received when the hands and eyes are otherwise engaged
- those with mobility impairment do not need to accommodate their physical position or location.

These benefits are likely to assume increasing importance as the number of older people (who tend to have an increased reliance on them) continues to grow [ONS, 2011]—a point not lost on the UK Government, which had identified spoken reminder systems as one way of promoting and supporting independent living in one's own home through technology, something it had identified as a key priority over the coming years [Department of Health, 2006]. Clearly, many of these benefits would be of interest to the wider population when receiving reminders or alarms.

Early work by Lines and Hone indicates that 'speech outputs appear to be a promising mode for interactive domestic alarm systems output' [Lines and Hone, 2003b]. A small-scale study in participants' own homes by Boman [2009] providing spoken reminders was successful in improving the ability to remember to perform activities and, therefore, quality of life for four of the five participants. In one of the few studies to compare speech with other auditory reminders [McGee-Lennon et al., 2007], speech significantly outperformed pager beeps and earcons [Blattner et al., 1989].

The results of these studies should come as no great surprise, given that speech is the primary means of communication for humans. The difficulty with spoken reminders is that, almost by definition, the receiver will be engaged in another task or, at the very least, will not be focused on the reminder. In the home environment (and, often, outside it), many of these other tasks will occur in the presence of noise; for example, talking with friends, listening to the radio, watching television, or vacuuming. As anyone who has tried to hold a conversation in a noisy environment will be aware, such background noise interferes with speech perception, making it difficult to hear and understand the other person.

7.4 Conclusion

The methodologies introduced and implemented in the preceding chapters offer an approach to assessing synthetic speech intelligibility that is grounded in the real world, using real-life stimuli in ecologically-valid noises and reverberation. So, rather than

the culmination of the work presented in this thesis, this chapter symbolizes the establishment of a starting point for carrying out further work.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Alamsaputra, D. M., Kohnert, K. J., Munson, B., and Reichle, J. (2006). Synthesized speech intelligibility among native speakers and non-native speakers of English. *Augmentative and Alternative Communication*, 22(4):258–268.

Aldrich, F. (1998). Pager messages as self reminders: A case study of their use in memory impairment. *Personal and Ubiquitous Computing*, 2(1):1–10.

Allen, J., Hunnicutt, M. S., and Klatt, D., editors (1987). *From text to speech: The MITalk system*. Cambridge Studies in Speech Science and Communication. Cambridge University Press, Cambridge, UK.

Allen, J. B. (2005). *Articulation and Intelligibility*. Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool Publishers, USA.

ANSI (2007). Methods for the calculation of the speech intelligibility index.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press, Cambridge.

Baker, J. (1975). The DRAGON system—an overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23.

Bates, D. and Maechler, M. (2009). lme4: Linear mixed-effects models using S4 classes.

Bench, J., Kowal, Å., and Bamford, J. (1979). The BKB (Bamford–Kowal–Bench) sentence lists for partially-hearing children. *British Journal of Audiology*, 13(3):108–112.

Bennett, C. L. (2005). Large scale evaluation of corpus-based synthesizers: Results and lessons from the Blizzard Challenge 2005. In *Ninth European Conference on Speech Communication and Technology*. ISCA.

Bennett, C. L. and Black, A. W. (2006). The Blizzard Challenge 2006. In *Proceedings of Blizzard Challenge Workshop, 2006*.

Benoît, C., Grice, M., and 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.

Benoît, C. and Pols, L. C. W. (1992). On the assessment of synthetic speech. In Bailly, G. and Benoît, C., editors, *Talking Machines: Theories, Models and Designs*, pages 435–442. Elsevier Science Publishers B. V., North-Holland, Amsterdam.

Bidelman, G. M. and Krishnan, A. (2010). Effects of reverberation on brainstem representation of speech in musicians and non-musicians. *Brain Research*, 1355:112–125.

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., and Rzeczkowski, C. (1984). Standardization of a test of speech perception in noise. *Journal of Speech, Language, and Hearing Research*, 27(1):32–48.

Bistafa, S. R. and Bradley, J. S. (2000). Reverberation time and maximum background-noise level for classrooms from a comparative study of speech intelligibility metrics. *The Journal of the Acoustical Society of America*, 107(2):861–875.

Black, A. W., Bunnell, H. T., Dou, Y., Muthukumar, P. K., Metze, F., Perry, D., Polzehl, T., Prahallad, K., Steidl, S., and Vaughn, C. (2012). Articulatory features for expressive speech synthesis. In *Proceedings of ICASSP 2012*, pages 4005–4008.

Black, A. W. and Taylor, P. (1994). CHATR: A generic speech synthesis system. In *Proceedings of the 15th Conference on Computational Linguistics*, pages 983–986, Morristown, New Jersey, USA. Association for Computational Linguistics.

Black, A. W. and Tokuda, K. (2005). The Blizzard Challenge — 2005: Evaluating corpus-based speech synthesis on common datasets. In *Interspeech 2005: 6th Annual Conference of the International Speech Communication Association*, pages 77–80.

Blattner, M. M., Sumikawa, D. A., and Greenberg, R. M. (1989). Earcons and icons: Their structure and common design principles. *SIGCHI Bulletin*, 21(1):123–124.

Boman, I.-L. (2009). *New technology and everyday functioning at home for persons with cognitive impairments after acquired brain injury*. PhD thesis, Karolinska Institutet, Stockholm, Sweden.

Boothroyd, A. (2005). Modeling the effects of room acoustics on speech reception and perception. In Crandell, J., Smaldino, J., and Flexer, C., editors, *Sound Field Amplification: Applications to Speech Perception and Classroom Acoustics*, pages 23–48. Thomson Delmar Learning, New York, 2nd edition.

Brand, T. (2009). Speech intelligibility. In Havelock, D., Kuwano, S., and Vorländer, M., editors, *Handbook of Signal Processing in Acoustics*, chapter 13, pages 197–204. Springer New York.

British Society of Audiology (2004). Recommended procedure: Pure tone air and bone conduction threshold audiometry with and without masking and determination of uncomfortable loudness levels. Available at: `http://www.thebsa.org.uk/docs/RecPro/PTA.pdf`.

Bunnell, H. T. and Lilley, J. (2007). Analysis methods for assessing TTS intelligibility. In *Proceedings of 6th Speech Synthesis Workshop (SSW6)*, pages 374–379.

Callison-Burch, C. and Dredze, M. (2010). Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12.

Campbell, G. A. (1910). XII. telephonic intelligibility. *Philosophical Magazine Series 6*, 19(109):152–159.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, L., Post, W., Reidsma, D., and Wellner, P. (2006). The AMI meeting corpus: A pre-announcement. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 3869 LNCS, pages 28–39.

Cernak, M. (2006). Unit selection speech synthesis in noise. In *2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings*, volume 1.

Chisholm, T. H., Willott, J. F., and Lister, J. J. (2003). The aging auditory system: Anatomic and physiologic changes and implications for rehabilitation. *International Journal of Audiology*, 42:2S3–2S10.

Christensen, H., Barker, J., Ma, N., and Green, P. (2010). The CHiME corpus: a resource and a challenge for computational hearing in multisource environments. In *Proceedings of Interspeech 2010*, pages 1918–1921.

Clark, R., Richmond, K., Strom, V., and King, S. (2006). Multisyn voice for the Blizzard Challenge 2006. In *Proceedings of Blizzard Challenge Workshop, 2006*.

Clark, R. A. J., Richmond, K., and King, S. (2007). Multisyn: Open-domain unit selection for the Festival speech synthesis system. *Speech Communication*, 49(4):317–330.

Cooke, M. (2003). Glimpsing speech. *Journal of Phonetics*, 31(3):579–584.

Cooke, M., Mayo, C., Valentini Botinhão, C., Stylianou, Y., Sauert, B., and Tang, Y. (2013). Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication*, 55(4):572–585.

Corso, J. F. (1959). Age and sex differences in pure-tone thresholds. *The Journal of the Acoustical Society of America*, 31(4):498–507.

Creer, S., Green, P., Cunningham, S., and Yamagishi, J. (2009). Building personalised synthesised voices for individuals with dysarthria using the HTS toolkit. In Mullennix, J. W. and Stern, S. E., editors, *Computer Synthesized Speech Technologies: Tools for Aiding Impairment*. IGI Global, 1st edition.

Department for Communities and Local Government (2012). English housing survey HOMES 2010. Available at: `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/6748/2173483.pdf`.

Department of Health (2006). Our health, our care, our say: A new direction for community services. Available at: `http://www.official-documents.gov.uk/document/cm67/6737/6737.pdf`.

Drager, K. D. R. and Reichle, J. (2010). CSS and children: Research results and future directions. In *Computer Synthesized Speech Technologies: Tools for Aiding Impairment*, chapter 8, pages 130–147. IGI Global.

Drager, K. D. R. and Reichle, J. E. (2001). 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(5):1052–1057.

Dreschler, W. A., Verschuure, H., Ludvigsen, C., and Westermann, S. (2001). ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. International Collegium for Rehabilitative Audiology. *Audiology*, 40(3):148.

Drullman, R. and Bronkhorst, A. W. (2000). Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *The Journal of the Acoustical Society of America*, 107(4):2224–2235.

Dutoit, T. and Leich, H. (1993). MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication*, 13(3–4):435–440.

Egan, J. P. (1948). Articulation testing methods. *Laryngoscope*, 58(9):991–995.

Fairbanks, G. (1958). Test of phonemic differentiation: The Rhyme Test. *The Journal of the Acoustical Society of America*, 30(7):596–600.

Falaschi, A., Giustiniani, M., and Verola, M. (1989). A hidden Markov model approach to speech synthesis. In *First European Conference on Speech Communication and Technology*. ISCA.

Fant, G. (2005). *Speech Acoustics and Phonetics: Selected Writings*, volume 24 of *Text, Speech and Language Technology*. Springer Netherlands.

Festen, J. M. and Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America*, 88(4):1725–1736.

Finney, D. (1952). *Probit analysis: A statistical treatment of the sigmoid response curve*. Cambridge University Press, Cambridge, 2 edition.

Fletcher, H. (1940). Auditory patterns. *Reviews of Modern Physics*, 12(1):47–65.

Fletcher, H. and Steinberg, J. C. (1930). Articulation testing methods. *The Journal of the Acoustical Society of America*, 1(2B):17–21.

Fraser, M. and King, S. (2007). The Blizzard Challenge 2007. In *Proceedings of Blizzard Challenge Workshop, 2007*.

French, N. R. and Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *The Journal of the Acoustical Society of America*, 19(1):90–119.

Friedman, S. (1983). DEC device turns text into speech. *MIS Week*, 4(51):1.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27(15):2865–2873.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

George, E. L. J., Festen, J. M., and Houtgast, T. (2008). The combined effects of reverberation and nonstationary noise on sentence intelligibility. *The Journal of the Acoustical Society of America*, 124(2):1269–1277.

Greene, B. G. (1986). Perception of synthetic speech by nonnative speakers of English. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, 30(13):1340–1343.

Hagerman, B. (1982). Sentences for testing speech intelligibility in noise. *Scandinavian Audiology*, 11(2):79–87.

Haigh, K. Z., Kiff, L. M., Myers, J., and Krichbaum, K. (2006). The Independent Lifestyle Assistant. *Assistive Technology*, 18:87–106.

Holmes, J. N., Mattingly, I. G., and Shearme, J. N. (1964). Speech synthesis by rule. *Language and Speech*, 7(3):127–143.

Hothorn, T., Hornik, K., van de Wiel, M., and Zeileis, A. (2008). Implementing a class of permutation tests: The coin package. *Journal of Statistical Software*, 28(8):1–23.

House, A. S., Williams, C. E., Hecker, M. H. L., and Kryter, K. D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *The Journal of the Acoustical Society of America*, 37(1):158–166.

Howell, P. and Kadi-Hanifi, K. (1991). Comparison of prosodic properties between read and spontaneous speech material. *Speech Communication*, 10(2):163–169.

Humes, L. E., Nelson, K. J., and Pisoni, D. B. (1991). Recognition of synthetic speech by hearing-impaired elderly listeners. *Journal of Speech and Hearing Research*, 34(5):1180–1184.

Hunt, A. and Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *ICASSP-96*, volume 1, pages 373–376, Atlanta, Georgia.

IEEE (1969). IEEE recommended practice for speech quality measurements. *Audio and Electroacoustics, IEEE Transactions on*, 17(3):225–246.

Ipeirotis, P. G. (2010). Demographics of Mechanical Turk. Technical report, New York University.

ITU (1993). Objective measurement of active speech level ITU-T recommendation P.56.

Jerger, J., Chmiel, R., Stach, B., and Spretnjak, M. (1993). Gender affects audiometric shape in presbyacusis. *Journal of the American Academy of Audiology*, 4(1):42–49.

Jeub, M., Schafer, M., and Vary, P. (2009). A binaural room impulse response database for the evaluation of dereverberation algorithms. In *Digital Signal Processing, 2009 16th International Conference on*, pages 1–5.

Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing*. Pearson Education, Upper Saddle River, NJ, USA, 2nd edition.

Kalikow, D. N., Stevens, K. N., and Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, 61(5):1337–1351.

Kangas, K. A. and Allen, G. D. (1990). Intelligibility of synthetic speech for normal-hearing and hearing-impaired listeners. *Journal of Speech and Hearing Disorders*, 55(4):751–755.

Karaiskos, V., King, S., Clark, R. A. J., and Mayo, C. (2008). The Blizzard Challenge 2008. In *Proceedings of Blizzard Challenge Workshop, 2008*.

Katz, J., Medwetsky, L., Burkard, R., and Hood, L. (2009). *Handbook of Clinical Audiology*. Lippincott Williams & Wilkins, Philadelphia, Pennsylvania USA, 6th edition.

Kewley-Port, D., Burkle, T. Z., and Lee, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 122(4):2365–2375.

Killion, M. C., Niquette, P. a., Gudmundsen, G. I., Revit, L. J., and Banerjee, S. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 116(October 2004):2395–2405.

Killion, M. C. and Villchur, E. (1993). Kessler was right—partly: But SIN test shows some aids improve hearing in noise. *The Hearing Journal*, 46(9):31–35.

King, S. and Karaiskos, V. (2009). The Blizzard Challenge 2009. In *Proceedings of Blizzard Challenge Workshop 2009*.

King, S. and Karaiskos, V. (2010). The Blizzard Challenge 2010. In *Proceedings of Blizzard Challenge Workshop 2010*.

King, S. and Karaiskos, V. (2011). The Blizzard Challenge 2011. In *Proceedings of Blizzard Challenge Workshop 2011*.

King, S. and Karaiskos, V. (2012). The Blizzard Challenge 2012. In *Proceedings of Blizzard Challenge Workshop 2012*.

Kittur, A., Chi, E. H., and Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing*, pages 453–456. ACM.

Kollmeier, B. and Wesselkamp, M. (1997). Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *The Journal of the Acoustical Society of America*, 102(4):2412–2421.

Koul, R. K. and Allen, G. D. (1993). Segmental intelligibility and speech interference thresholds of high-quality synthetic speech in presence of noise. *Journal of Speech and Hearing Research*, 36(4):790–799.

Lancaster, J. A., Robinson, G. S., and Casali, J. G. (2004). Comparison of two voice synthesis systems as to speech intelligibility in aircraft cockpit engine noise. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, pages 127–131.

Langner, B. and Black, A. (2004). An examination of speech in noise and its effect on understandability for natural and synthetic speech. Technical report, Language Technologies Institute, CMU, Pittsburgh Pennsylvania.

Langner, B. and Black, A. W. (2005). Using speech in noise to improve understandability for elderly listeners. In *Proceedings of ASRU, San Juan, Puerto Rico*, pages 392–396. Citeseer.

Laredo, J. and Gould, G. (2007). *Bach: The Six Sonatas for Violin and Harpsichord*. Columbia Masterworks.

Larm, P. and Hongisto, V. (2006). Experimental comparison between speech transmission index, rapid speech transmission index, and speech intelligibility index. *The Journal of the Acoustical Society of America*, 119(2):1106–1117.

Lehiste, I. and Peterson, G. E. (1959). Linguistic considerations in the study of speech intelligibility. *The Journal of the Acoustical Society of America*, 31(3):280–286.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, 49(2B):467–477.

Levitt, H. and Rabiner, L. R. (1967). Use of a sequential strategy in intelligibility testing. *The Journal of the Acoustical Society of America*, 42(3):609–612.

Licklider, J. C. R. and Miller, G. A. (1948). The intelligibility of interrupted speech. *The Journal of the Acoustical Society of America*, 20(4):593.

Lines, L. and Hone, K. S. (2002). Older adults' evaluations of speech output. In *5th International ACM SIGCAPH Conference on Assistive Technologies (ASSETS 2002)*, pages 170–177, New York. ACM.

Lines, L. and Hone, K. S. (2003a). Older adults' comprehension and evaluation of speech as alarm system output within the domestic environment. In *2nd International Conference on Universal Access in Human Computer Interaction, Crete, Greece*.

Lines, L. and Hone, K. S. (2003b). Older adults' comprehension of speech as interactive domestic alarm system output: A field study. In *Proceedings of HCI International 2003*, Crete, Greece. Lawrence Erlbaum Associates, Inc.

Ljolje, A. and Fallside, F. (1986). Synthesis of natural sounding pitch contours in isolated utterances using hidden Markov models. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 34(5):1074–1080.

Loizou, P. C. (2007). *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*. CRC.

Lombard, E. (1911). Le signe de l'élévation de la voix. *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, 37(101–119).

Mack, M. (1988). Sentence processing by non-native speakers of English: Evidence from the perception of natural and computer-generated anomalous L2 sentences. *Journal of Neurolinguistics*, 3(2):293–316.

Marge, M., Banerjee, S., and Rudnicky, A. I. (2010). Using the Amazon Mechanical Turk for transcription of spoken language. In *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 5270–5273. IEEE.

McArdle, R. and Wilson, R. H. (2009). Speech perception in noise: The basics. *Perspectives on Hearing and Hearing Disorders: Research and Diagnostics*, 13(1):4–13.

McCarty, C. Q. and Surprenant, A. (2006). Older adult's identification and memory of synthetic and natural speech in noise. *The Journal of the Acoustical Society of America*, 120(5):3347.

McGee-Lennon, M., Wolters, M., and McBryan, T. (2007). Audio reminders in the home environment. In *Proceedings of International Conference on Auditory Display (ICAD), Montreal, Canada*.

Meyer, R., Brand, T., and Kollmeier, B. (2007). Predicting speech intelligibility in fluctuating noise. In *8th EFAS Congress/10th Congress of the German Society of Audiology*, Heidelberg, Germany.

Miller, G. A. and Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, 2(3):217–228.

Miller, G. A. and Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, 27(2):338–352.

Mirenda, P. and Beukelman, D. R. (1987). A comparison of speech synthesis intelligibility with listeners from three age groups. *Augmentative and Alternative Communication*, 3(3):120–128.

Möller, S. (2004). Telephone transmission impact on synthesized speech: Quality assessment and prediction. *Acta Acustica United with Acustica*, 90(1):121–136.

Morgan, D. E., Kamm, C. A., and Velde, T. M. (1981). Form equivalence of the speech perception in noise (SPIN) test. *The Journal of the Acoustical Society of America*, 69(6):1791–1798.

Morrison, H. B. and Casali, J. G. (1997). Intelligibility of synthesized voice messages in commercial truck cab noise for normal-hearing and hearing-impaired listeners. *International Journal of Speech Technology*, 2(1):33–44.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6):453–467.

Nabelek, A. K. and Mason, D. (1981). Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms. *Journal of Speech and Hearing Research*, 24(3):375–383.

Nabelek, A. K. and Robinson, P. K. (1982). Monaural and binaural speech-perception in reverberation for listeners of various ages. *The Journal of the Acoustical Society of America*, 71(5):1242–1248.

Newman, C. W., Weinstein, B. E., Jacobson, G. P., and Hug, G. A. (1991). Test-retest reliability of the Hearing Handicap Inventory for Adults. *Ear and Hearing*, 12(5):355.

Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, 95(2):1085–1099.

Niquette, P., Arcaroli, J., Revit, L., Parkinson, A., Staller, S., Skinner, M., and Killion, M. (2003). Development of the BKB-SIN test. In *Annual meeting of the American Auditory Society, Scottsdale, AZ*.

Nixon, C. W., Stephenson, M. S., McKinley, R. L., Fisher, V. G., and Jacobs, M. J. (1990). Intelligibility of digital speech masked by noise: Normal hearing and hearing impaired listeners. Technical report, Harry G. Armstrong Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio, USA.

Novotney, S. and Callison-Burch, C. (2010). Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 207–215.

Noyes, J. M., Hellier, E., and Edworthy, J. (2006). Speech warnings: A review. *Theoretical Issues in Ergonomics Science*, 7(6):551–571.

Nugent, C. D., Finlay, D. D., Fiorini, P., Tsumaki, Y., and Prassler, E. (2008). Editorial home automation as a means of independent living. *Automation Science and Engineering, IEEE Transactions on*, 5(1):1–9.

Nye, P. W. and Gaitenby, J. H. (1974). The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. Technical report, Haskins Laboratories Status Report on Speech Research, SR-37.

OED Online (2013). crowdsourcing, n. Available at: http://www.oed.com/view/Entry/376403.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, 9(1):97–113.

Olive, J. (1977). Rule synthesis of speech from dyadic units. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '77.*, volume 2, pages 568–570.

ONS (2011). National population projections, 2010-based statistical bulletin. Technical Report October, Office for National Statistics, London.

Payton, K. L., Uchanski, R. M., and Braida, L. D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *The Journal of the Acoustical Society of America*, 95(3):1581–1592.

Perry, M., Dowdall, A., Lines, L., and Hone, K. (2004). Multimodal and ubiquitous computing systems: Supporting independent-living older users. *Information Technology in Biomedicine, IEEE Transactions on*, 8(3):258–270.

Pisoni, D. B. (1987). Some measures of intelligibility and comprehension. In *From text to speech: The MITalk system*, chapter 13, pages 151–171. Cambridge University Press.

Pisoni, D. B. and Koen, E. (1982). Some comparisons of intelligibility of synthetic and natural speech at different speech-to-noise ratios. *The Journal of the Acoustical Society of America*, 71:S94.

Plomp, R. and Mimpen, A. M. (1979). Speech-reception threshold for sentences as a function of age and noise level. *The Journal of the Acoustical Society of America*, 66(5):1333–1342.

R Development Core Team (2009). R: A language and environment for statistical computing.

Raitio, T., Takanen, M., Santala, O., Suni, A., Vainio, M., and Alku, P. (2012). On measuring the intelligibility of synthetic speech in noise — do we need a realistic noise environment? In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4025–4028. IEEE.

Raynor, D. K., Booth, T. G., and Blenkinsopp, A. (1993). Effects of computer generated reminder charts on patients' compliance with drug regimens. *BMJ*, 306(6886):1158–1161.

Reynolds, M., Bond, Z. S., and Fucci, D. (1996). Synthetic speech intelligibility: Comparison of native and non-native speakers of English. *Augmentative and Alternative Communication*, 12(1):32–36.

Rhebergen, K. S. and Versfeld, N. J. (2005). A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 117:2181.

RNIB (2010). RNIB online shop. Available at: `http://onlineshop.rnib.org.uk`.

Roring, R. W., Hines, F. G., and Charness, N. (2007). Age differences in identifying words in synthetic speech. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 49:25–31.

Rothauser, E. H., Urbanek, G. E., and Pachl, W. P. (1968). Isopreference method for speech evaluation. *The Journal of the Acoustical Society of America*, 44(2):408–418.

Roweis, S. (2003). Factorial models and refiltering for speech separation and denoising. *Proceedings of Eurospeech 2003*, pages 1009–1012.

Sabine, W. C. (1923). *Collected Papers On Acoustics*. Harvard University Press, Cambridge.

Sagisaka, Y., Kaiki, N., Iwahashi, N., and K., M. (1992). ATR v-Talk speech synthesis system. In *International Conference on Spoken Language Systems*, pages 483–486.

Shneiderman, B. and Plaisant, C. (2004). *Designing the User Interface: Strategies for Effective Human-Computer Interaction (4th Edition)*. Pearson Addison Wesley.

Smith, C. (1979). Talker variance and phonetic feature variance in diagnostic intelligibility scores for digital voice communications processors. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79*, volume 4, pages 456–459.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast — but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263.

Steeneken, H. J. M. and Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *The Journal of the Acoustical Society of America*, 67(1):318–326.

Summers, W. V. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, 84(3):917–928.

Tachibana, M., Yamagishi, J., Masuko, T., and Kobayashi, T. (2005). Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. *IEICE Transactions on Information and Systems*, 88(11):2484–2491.

Taylor, P. (2009). *Text-to-Speech Synthesis*. Cambridge University Press, Cambridge, UK.

Toda, T. and Tokuda, K. (2005). Speech parameter generation algorithm considering global variance for HMM-based speech synthesis. In *Ninth European Conference on Speech Communication and Technology*, pages 2801–2804. ISCA.

Tokuda, K., Masuko, T., Yamada, T., Kobayashi, T., and Imai, S. (1995). An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features. In *Fourth European Conference on Speech Communication and Technology*, pages 757–760. ISCA.

Valentini Botinhão, C., Yamagishi, J., and King, S. (2011a). Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise? In *Interspeech 2011*, pages 1837–1840.

Valentini Botinhão, C., Yamagishi, J., and King, S. (2011b). Evaluation of objective measures for intelligibility prediction of HMM-based synthetic speech in noise. In *ICASSP 2011*, pages 5112–5115.

Valentini Botinhão, C., Yamagishi, J., King, S., and Maia, R. (2013). Intelligibility enhancement of HMM-generated speech in additive noise by modifying mel cepstral coefficients to increase the glimpse proportion. *Computer Speech & Language*.

van Leeuwen, D. A. and van Balken, J. (2005). Evaluation of speech synthesis systems using the speech reception threshold methodology. In *New Directions for Improving Audio Effectiveness*, pages 9.1–9.6.

van Wijngaarden, S. J., Steeneken, H. J. M., and Houtgast, T. (2002). Quantifying the intelligibility of speech in noise for non-native listeners. *The Journal of the Acoustical Society of America*, 111(4):1906–1916.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition.

Venkatagiri, H. S. (2003). Segmental intelligibility of four currently used text-to-speech synthesis methods. *The Journal of the Acoustical Society of America*, 113(4):2095–2104.

Voiers, W. D. (1983). Evaluating processed speech using the Diagnostic Rhyme Test. *Speech Technology*, 1(4):30–39.

von Kempelen, W. (1791). *Mechanismus der menschlichen Sprache und Beschreibung einer sprechenden Machine*. J B Degen.

Wagener, K. (2009). Multilingual speech tests in several European countries. Technical report, Hörzentrum Oldenburg.

Wagener, K. C., Brand, T., and Kollmeier, B. (2007). International cross-validation of sentence intelligibility tests. In *8th EFAS Congress/10th Congress of the German Society of Audiology*, pages 1–3.

Watts, O., Yamagishi, J., and King, S. (2010). Letter-based speech synthesis. In *Proceedings of Speech Synthesis Workshop 2010*, pages 317–322.

Whitman, F. P. (1915). On the acoustics of the chapel of adelbert college. *Science*, pages 191–193.

Wilson, R. H. (2003). Development of a speech-in-multitalker-babble paradigm to assess word-recognition performance. *Journal of the American Academy of Audiology*, 14(9):453–470.

Wilson, R. H., McArdle, R. A., and Smith, S. L. (2007). An evaluation of the BKB-SIN, HINT, QuickSIN, and WIN materials on listeners with normal hearing and listeners with hearing loss. *Journal of Speech, Language, and Hearing Research*, 50(4):844–856.

Winters, S. J. and Pisoni, D. B. (2006). Speech synthesis, perception and comprehension of. In Brown, K., editor, *Encyclopedia of Language & Linguistics*, pages 31–49. Elsevier, Oxford.

Wolters, M., Campbell, P., DePlacido, C., Liddell, A., and Owens, D. (2007). Making speech synthesis more accessible to older people. In *Proceedings of Sixth ISCA Workshop on Speech Synthesis, Bonn, Germany*.

Wolters, M. K., Isaac, K. B., and Doherty, J. M. (2012). Hold that thought: Are spearcons less disruptive than spoken reminders? In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, pages 1745–1750.

Wolters, M. K., Isaac, K. B., and Renals, S. (2010). Evaluating speech synthesis intelligibility using Amazon Mechanical Turk. In *Proceedings of 7th Speech Synthesis Workshop (SSW7)*, pages 136–141.

Wolters, M. K., Johnson, C., Campbell, P. E., DePlacido, C. G., and McKinstry, B. (2014). Can older people remember medication reminders presented using synthetic speech? *Journal of the American Medical Informatics Association : JAMIA*, pages 1–6.

Wolters, M. K., Johnson, C., and Isaac, K. B. (2011). Can the Hearing Handicap Inventory for Adults be used as a screen for perception experiments? In *ICPhS XVII 2011*, Hong Kong.

Yamagishi, J., Nose, T., Zen, H., Ling, Z., Toda, T., Tokuda, K., King, S., and Renals, S. (2009). Robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, 17(6):1208–1230.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., and Tokuda, K. (2007). The HMM-based speech synthesis system (HTS) version 2.0. In *6th ISCA Workshop on Speech Synthesis*, pages 294–299.

Zen, H. and Toda, T. (2005). An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In *Ninth European Conference on Speech Communication and Technology*.

Zen, H., Tokuda, K., and Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064.

Zhou, F., Yang, H.-I., Álamo, J. M. R., Wong, J. S., and Chang, C. K. (2010). Mobile personal health care system for patients with diabetes. In Lee, Y., Bien, Z. Z., Mokhtari, M., Kim, J. T., Park, M., Kim, J., Lee, H., and Ibrahim, I. K., editors, *ICOST*, volume 6159 of *Lecture Notes in Computer Science*, pages 94–101. Springer.