Further Investigation of MDS as a Tool for Evaluation of Speech Quality of Synthesized Speech

7487663

Speech and Language Processing
School of Philosophy, Psychology & Language Sciences
University of Edinburgh

August 21st, 2009

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF THE DEGREE OF
MASTER OF SCIENCE

Abstract

The dissertation investigates MDS as a tool for the evaluation of the quality of synthesized speech. More specifically, it investigates the relations between Weighted Euclidean Distance Scaling and Simple Euclidean Distance Scaling, and how aggregating data affects the MDS configuration. It is investigated to what extent a subset of experimental participants and/or experimental stimuli are representative of a larger test set. For that purpose an experiment was conducted on the basis of a subset of stimuli used in the Blizzard Challenge 2008. Issues in the evaluation of Speech Synthesis are discussed and an overview of the basics of multi-dimensional scaling is given to an extent that allows comprehension of methods used in the application of Multi-dimensional scaling to speech synthesis evaluation. Based on the experimental findings, further experiments are suggested with the goal in mind that testing procedures can be optimized to such an extent that the number of experimental participants can be drastically reduced.

Acknowledgements

I would like to thank Rob Clark for his patience, support, and soothing calm throughout this MSc year in general, and during the supervison of this dissertation in particular. A very heartfelt thank you also goes out to my colleagues who were always willing to help out in times of need of coffee breaks, python script error spotting, a good rant, and, most importantly, a good laugh. I am particularly indebted to Oliver, who has -again- proved to be a life-saver in all areas possibly conceivable. Furthermore, my thanks go to Fletch, who always obliged to smile and nod happily, when I told her about my small daily triumphs in tackling this script or solving that bug, even though she had not the slightest idea what I was actually talking about; and thank you for engaging me in completely unrelated conversation afterwards to get my mind off it, when I found out that my triumphs actually were none, but just another bug in a script. Last, but definitely not least, thank you, mum and dad, for supporting me throughout my studies, keeping me free of financial as well as logistic worries, as well as keeping up my morals. I could not have done it without your unconditional support.

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

"Come and play with me," proposed the little prince, "I am so unhappy."
"I cannot play with you," the fox said, "I am not tamed."[...]
The fox gazed at the little prince, for a long time. "Please–tame me!" he said.
"I want to, very much," the little prince replied. "But I have not much time.
I have friends to discover, and a great many things to understand."
"One only understands the things that one tames," said the fox.
Antoine de de Saint-Exupry, *The Little Prince*

Evaluation plays a vital role in advancing the state-of-the-art in text-to-speech synthesis (TTS) research. However, this is easier said than done: developing an accurate evaluation paradigm is far from trivial. (Podsiadlo 2007) Multidimensional scaling (MDS) has recently been discovered as a powerful tool for the evaluation of the quality of speech generated by TTS. (Mayo et al. 2005, Podsiadlo 2007, Hall 2001) Its big disadvantage is, however, that when it is applied to a large-scale data set, it relies on subjective evaluation of a vast number of people. In order to add a new stimulus point to an already existing MDS matrix, testing of that stimulus against all previously fitted stimuli has to occur. This is incredibly costly and time-consuming. "The development of a good automatic metric for synthesis evaluation, one that would eliminate the need for expensive and time-consuming human listening experiments, remains an open and exciting research question." (Jurafsky & Martin 2008, p. 314) We are far from there, yet, and the current ambition is to minimize the amount of human testing that is needed for fitting new data into an already existing object space generated by MDS. However, in order to do so, we first need to learn more about the tool of MDS applied to perceptual judgments of TTS systems. We need to compare and contrast different MDS functions, and investigate the relation between the single listener's judgment and the overall result. For that purpose, we also need to

know, whether some listeners are "more average than others". In this paper, the influence the variable of native language exerts on speech quality evaluation will be investigated.

To gain better understanding of these issues raised, a listening experiment has been conducted. It is based on a subset of the data submitted in the Blizzard Challenge 2008 and investigates the effects of different MDS settings on the analysis of the data, as well as compare the results smaller test samples of participants and stimuli, respectively, yield in comparison to larger scale evaluations.

1.1 Outline

Chapter 2 will be an overview of TTS system evaluation: TTS system evaluation will be described in terms of criteria to be assessed, measures of their quality, and methods to obtain named measures. A few commonly used methods for the assessment of speech intelligibility and speech quality will be named and their strengths and weaknesses will be discussed. It is vital to understand how the evaluation of speech quality differs from that of speech intelligibility and the difficulties the abstract and often vague measures pose for evaluation. Then a few projects for the comparison of speech quality across TTS systems will be introduced, including the *Blizzard Challenge*, which is of particular interest to this paper, as the experiment described in the practical part relies on data from the Blizzard Challenge 2008. Chapter 3 describes the basics of MDS and contrasts different techniques along the variables of input data, functions, and models employed. The powers and pitfalls of MDS techniques are described, and the interpretation of an MDS output graph will be explained. The focus is then narrowed down to investigate MDS applied to speech quality evaluation of synthesized speech, by giving an overview of the current state of research in the field. Based on that, Chapter 4 addresses a few selected research questions that have been left unanswered by previous research in the field. To jab at these unknown areas, an experiment was devised; The experimental design and methodological choices are explained and put into relation to the research questions. Chapter 5 then proceeds to the analysis and discussion of the experimental data.

CHAPTER 2

Evaluation of TTS systems

It was easier to know it than to explain why I know it. If you were asked to prove that two and two made four, you might find some difficulty, and yet you are quite sure of the fact.
Sir Arthur Conan Doyle, *A Study in Scarlet*

Evaluation of TTS systems aids developers to increase their understanding of a system's strengths and weaknesses, as well as to make decisions about parts of a system in most vital need of improvement. It can determine what expectations a user has of the system, and can even give linguists insights into human speech processing. (Campbell 2005)

To arrive at that, three components have to be defined when evaluating a system (Gibbon et al. 2000):

1. **Criterion**: what characteristic or quality is evaluated?
2. **Measure**: in which property does the quality under investigation manifest itself?
3. **Method**: given a system, how are the measures observed?

2.1 Criteria

2.1.1 Global performance evaluation

Research in automatic speech recognition has greatly profited from comparisons against benchmarks of global performance. To evaluate a new system, it is simply presented with a specific standardized test set which has not been used for training of that system. Accuracy is then commonly used to report a system's

performance. (Young et al. 1995, p. 178) This makes evaluation co-text and context independent, i.e. absolute. Obviously, such an absolute measure would also be desirable for global performance evaluation of TTS. The application of this approach is not as straightforward, though: While it is easily said that the best recognition system is that one which achieves the highest accuracy of recognition, people's ideas about the best speech synthesis system are less unanimous. Is it the systems which can be most easily understood? Or is it the system which sounds most natural?

Also, a drawback of performance evaluation is that it can only account for the acceptance of a system, but it has no diagnostic value whatsoever; in other words: when compiling such a test, we will find out which system users like best, but we do not know why. (Pols 1998)

2.1.2 Modular diagnostic evaluation

Modular diagnostic evaluation assesses the performance of a module of a TTS system, such as "text pre-processing, grapheme-to-phoneme conversion, phrasing, accentual (focus), phoneme intelligibility, word and (proper) name intelligibility [... by measuring] performance with ambiguous sentences, [in] comprehension tests, and psycho-linguistic tests such as lexical decision and word recall". (Pols 1998, p. 501) In 1995, Pols remarked on the scarcity of benchmarks and standardizations in modular diagnostic evaluation, as well as the great lack of proper tests concerning prosody, speaker style, and emotion characteristics, which he at least partially appoints to the fact that TTS systems have not mastered them, yet. (Pols 1998, p. 501) However, since then research has not stood still: Standardized tests of phoneme and word intelligibility are common practice (cf. section 2.3.1), and the Blizzard Challenge (cf. section 2.4.2), which addresses several of the former issues, has been called into existence. TTS systems have improved drastically, and state-of-the-art systems are very intelligible. Prosody research for TTS has advanced (e.g. de Cheveigne & Kawahara (2002), Malfrère et al. (1998), Raux & Black (2003)) and research in speaker characteristics and synthesizing emotional speech is one of the current hot topics (e.g. Montero et al. (1998), Bulut et al. (2002), Turk et al. (2005)). It is exactly these suprasegmental levels that account for the difference between decent and outstanding systems; yet, it is just these differences, that are hard to test. (Mayo et al. 2005)

2.2 Measures

2.2.1 Speech intelligibility measures

Intelligibility is defined as "the ability of a human listener to correctly interpret the words and meaning of the synthesized utterance". (Jurafsky & Martin 2008, p. 314)

2.2.2 Speech quality measures

We determined above that speech cannot be right or wrong as a whole. However, when listening to TTS, the output sometimes seems "not to sound quite right", while being intelligible. This is due to some aspect speech quality. Speech quality is defined as "an abstract measure of the naturalness, fluency, or clarity of the speech". (Jurafsky & Martin 2008, p. 314)

2.3 Methods

2.3.1 Subjective evaluation of speech intelligibility

Intelligibility tests are the part of TTS evaluation that probably comes closest to being analogous with the accuracy measure benchmark in the evaluation of automatic speech recognition, as all these results are absolute. The three probably most common intelligibility tests are the *Diagnostic Rhyme Test (DRT)*, the *Modified Rhyme Test (MRT)*, and the *Semantically unpredictable sentences (SUS) test*, but there are more, most of which are variations of these three.

The DRT is a forced decision task, constructed from a subset of a set of 96 minimal pairs (i.e. words that differ in only one sound from each other), whose initial consonants differ in only one feature (e.g. $\pm$ voiced). Listeners have to indicate which of the two words they think they heard. (Voiers et al. 1975) The percentage of correctly identified words is used for comparison of systems. (Jurafsky & Martin 2008, p. 315)

In the MRT, experimental participants are presented with words, synthesized by a TTS system, each inside a carrier phrase *Now we will say.....again*. The words to be synthesized are taken from a set of 300 words. Each of these stimuli is part of a set of six words which are identical, with the exception of their initial or final consonants. Listeners must identify the word they heard within its set of

words. (House et al. 1963) The percentage of correctly identified words is used for comparison of systems.

The SUS test was created to allow a standardized measure for comparison (potentially across languages) of intelligibility of connected speech generated by TTS systems. The SUS task presents syntactically correct sentences of a simple structure that are semantically nonsensical, and participants are requested to transcribe these sentences. Due to the sentences' intact syntactic structure, listeners can determine each word's part of speech from the structure, while the sentences remain devoid of global meaning. This has the benefit that it can be ruled out that semantic context allows listeners to guess a word they otherwise would not recognize. The percentage of correctly transcribed sentences is used for comparison of systems. (Benoit & Grice 1996)

2.3.2 Subjective evaluation of speech quality

The section on speech intelligibility shows that exact measures of performance which also serve as a mean for comparison are already in place in that area. The tests can be applied individually, at different times, for different systems, in different contexts. This, however, is not at all true for the subjective evaluation of speech quality, yet, since the output generated cannot simply be measured according to its accuracy, as there is no such thing as "right speech", or "wrong speech". Consequently, to this day, global evaluation of TTS systems is only relative, i.e.the results obtained in these tests are only relative, i.e. they are only meaningful within the context of the systems tested. Further systems cannot simply be added at a later point without testing any new system against systems present in the initial test configuration.

Mean opinion scores (MOS) are a common method for gathering subjective judgments of speech quality. MOS are derived by analysis of untrained listeners' ratings of stimuli along a scale, usually from 1 (bad) to 5 (good). These scores are only valid within the context they were tested in. One sample whose quality differs drastically from that of the samples surrounding it is likely to be appointed more extreme scores than it would be within samples of similar quality. Furthermore, listeners differ, and these values are not absolute. (Hall 2001, p. 2167) As such, Taylor (2009) suggest to consider MOS tests as "ranking tests". (p. 537)

Also, since evaluation of speech quality is based on abstract categories, identifying distinct dimensions can be hard for experimental participants. Furthermore,

when asked to listen to one of these dimensions, more perceptually salient dimensions tend to influence their judgments. (Mayo et al. 2005, Taylor 2009) A study conducted by Alvarez & Huckvale (2002) found that judgments across dimensions were highly correlated, which, they suggested, indicates that listeners were not assessing clearly distinct dimensions of the systems under investigation. Hirst et al. (1998) and Vainio et al. (2002) have reported similar results. Hence, reliable results will only be obtained, when speech quality is tested as a whole, and listeners are not asked to make distinctions on a level lower than that.

As such,

> [t]he single composite judgment of quality provided by MOS testing is essential for acceptance testing, but it does not tell us *why* the quality is good or bad. (Hall 2001, p. 2168)

Multi-dimensional scaling has recently been discovered as a powerful tool for the evaluation of speech quality. Mayo et al. (2005) summarize the current state of affairs in research in MDS for TTS evaluation:

> Unfortunately, no one has examined the acoustic dimension weighting behaviour of listeners when rating synthetic speech. It is therefore unclear whether listeners are, for example, consistently more influenced in a speech synthesis rating task by segmental quality, or by appropriateness of intonation. The [ultimate] goal of the current line of research [in MDS as evaluation tool], therefore, is to determine the pattern of weights listeners give to available acoustic dimensions (both sub- and supra-segmental) when rating synthetic speech.(p.1)

More on the power and potential of MDS in general and as a tool for TTS evaluation will follow in chapters below.

2.3.3 Objective evaluation of speech quality

Objective evaluation provides a mathematical measure for the relation between two waveforms, thus it lends itself for the evaluation of cases in which the aim of a system is the reproduction of an original input waveform. By measuring signal-to-noise-ratio (SNR), i.e. "the average energy of the original speech waveform to the average energy in the error (or 'noise') signal representing the distortion introduced by the [TTS system's] coding algorithm", the original waveform and a

TTS system's output are compared. (Holmes 2001, p. 64f) Objective evaluation methods do not yet distinguish between acoustic features that are perceptually salient and those that are not. The goodness of an objective evaluation method is defined by its fit to human perception. (Holmes 2001, p. 63-65) And thus, the main problem of objective evaluation of speech is that results gained from distance measurements between synthesized speech and recorded natural speech do not necessarily match the results generated by human listeners in subjective evaluation tasks. (Taylor 2009, p. 538) Experiments conducted by Clark & Dusterhoff (1999) suggest that the relation between metric measurements of prosody in synthesized speech and human perception of those is a non-linear one. The exact nature of this relation has not yet been identified, though.

2.4 Comparing speech quality generated by different systems

As indicated above, the highly standardized nature of intelligibility tests renders them very suitable for comparisons across systems. The following section examines projects that (also) approached the more taxing task of arriving at comparisons of speech quality across systems.

2.4.1 ITU-T P.85

The International Telecommunication Union devised recommendation P.85, *A method for subjective performance assessment of the quality of speech output devices*, as a guideline of TTS system comparison, which should allow global performance evaluation as well as modular diagnostic evaluation. It is a series of listening tasks, during which it is first established that listeners understand the content, and then several MOS are collected.(P.85 1994) The weakness of the project lies in its method: The listeners do not arrive at assessing purely the dimension they are instructed to listen to, and scores for different dimensions are highly correlated.(Alvarez & Huckvale 2002, Sityaev et al. 2006)

2.4.2 The Blizzard Challenge

The Blizzard Challenge, which has been held annually since 2005, is a "research exercise"(Karaiskos et al. 2008) for TTS systems, which has borrowed the underlying idea of its design from ASR systems, namely that a common dataset is used for testing. All participants of the challenge build a voice from the same speech database. Hence, the effectiveness of different TTS techniques and the quality

of systems are directly comparable. (Black & Tokuda 2005, p. 77) This is important, as the one system's acceptance by listeners can vary depending on the voice used.

The evaluation consists of intelligibility testing by means of a SUS task, and the evaluation of speech quality by collecting similarity-difference judgments for MDS analysis, MOS of overall naturalness, and MOS rating the similarity of test sentences to reference sentences.Fraser & King (2007)

A vital point in the Blizzard Challenge is that every year two systems of the previous year are included in the evaluation as benchmarks. So even though the results of this large scale evaluation are not absolute, their scope is widened.

Taylor (2009) deems the Blizzard challenge a potential driving force in the rate of development in TTS research, since its format allows "performance differences between systems [... to] easily be seen, and the competitive nature of the evaluation program has been credited with driving forward progress'.(p. 539)

2.5 Summarizing the major issues in TTS evaluation

Currently, one of the biggest issues in subjective evaluation of speech quality is not ranking stimuli or systems, but explaining how these ranks were derived. Up to date, no system will be mistaken for a human speaker in all its output; systems are far from sounding perfectly natural, and various imperfections are aggregated. This is less trivial than it sounds; listeners have a good notion of a system's speech quality, and a system that does not sound right is easily detected. It is much harder, though, to explain why it does not sound right. Ultimately, an objective measure for TTS evaluation is the desirable research goal, but this will not be feasible until we have gained better understanding of listeners' perceptual behaviour in evaluating synthesized speech; in particular this means: knowing how different dimensions are weighted in human perceptual processing of synthesized speech. This could either be done by gaining insights into the human processing mechanisms, or by extracting patterns by applying machine learning to large corpora of data of subjective evaluations, collected in large scale TTS evaluations. Meanwhile, to bridge the gap between the current state of depending on a legion of participants in listening experiments, and the (distant) goal of fully objective evaluation of speech quality, the intermediate research goal is

to investigate if/how/where/under what circumstances the amount of testing needed in the evaluation of speech quality can be reduced.

CHAPTER 3

Multi-dimensional scaling (MDS)

Turn him to any cause of policy,
The Gordian Knot of it he will unloose,
Familiar as his garter.
William Shakespeare, *Henry V*

Multidimensional scaling (MDS) is "a family of models by means of which information contained in a set of data is represented by a set of points in a space."(Coxon et al. 1982, p. 1) This space is constructed in such a way that the metric distance between the points in it is analogous to the empirical distance observed in the data. An advantage of MDS is that it produces visual output, and by literally just *looking at* the data, the data and its structure it become more comprehensible. (Coxon et al. 1982)

Multidimensional scaling maps "proximities p_{ij} [...] into distances of an m-dimensional MDS configuration X [..., defined] by a *representation function* $f(p_{ij})$ that specifies how the proximities should be related to distances $d_{ij}(X)$." (Borg & Groenen 2005, p. 37) Hence every instant of MDS is characterized by three choices made by the researcher (Coxon et al. 1982):

- **the data** to be analyzed
- **the transformation** defining the information in the data that is to be represented in the solution.
- **the model** to interpret the data

3.1 The data

The empirical distance d_{ij} between two objects i and j can either be gathered directly, by collecting numerical or order estimates in an experiment, or it can be derived, by aggregating direct data. Coxon et al. (1982) recommends to always check for the existence of different subgroups within the population tested. Different behavioral patterns of different groups may cancel one another out in aggregated data and produce an average that is not representative of any of the groups. *Piecemeal distortion* is another problem: "If the data referring to a given unit or individual is complex, then 'local structure' (interrelationships within parts of the data) can be lost entirely when the components are aggregated".(Coxon et al. 1982, p. 15) It thus is advisable to first examine the behaviour of individual subjects before proceding to aggregating data.

3.2 The transformation

The representation function $p_{ij}(X)$ defines how proximities p_{ij} are transformed into distances $d_{ij}(X)$ of an MDS space X, i.e.

$$f : p_{ij} \rightarrow d_{ij}(X) \tag{3.1}$$

The transformation f specifies the *MDS model*. MDS models are based on the stipulation that they are *exact* definitions of how an empirical distance d_{ij} is transformed into a proximity p_{ij} in the MDS solution space, i.e. after a transformation f, proximities p_{ij} equal lengths of edges d_{ij} between points i and j in a configuration X:

$$f : p_{ij} = d_{ij}(X) \tag{3.2}$$

Equation 3.2, however, is not quite accurate, as empirical measurements are never 100 per cent exact, but are to some extent distorted by noise, i.e. error. Thus, to be precise, we must not assume that the p_{ij} *equals* $d_{ij}(X)$, but that it is an approximation: (Borg & Groenen 2005, p. 41)

$$f : p_{ij} \approx d_{ij}(X) \tag{3.3}$$

Statistical computer programmes use an initial configuration $f(p_{ij})$, which then is adapted iteratively to approach $d_{ij}(X)$ as closely as possible.

The slight imprecision caused by the discrepancy between $f(p_{ij})$ and $d_{ij}(X)$ is measured by a *Stress function*. (Borg & Groenen 2005, p. 41)

3.2.1 The stress function

Stress is a measure of error. A squared error of representation is defined as

$$e_{ij}^2 = \left[f\left(p_{ij}\right) - d_{ij}\left(X\right)\right]^2 \tag{3.4}$$

This error, as in equation 3.4, is summed over all edges of the MDS representation to equal what is called *Raw Stress*

$$\sigma_r = \sigma_r\left(X\right) = \sum_{ij}\left[f\left(p_{ij}\right) - d_{ij}\left(X\right)\right]^2 \tag{3.5}$$

Raw stress is not particularly informative, though, as it is highly dependent on the configuration and measure of the data. To filter out this effect, σ_r can be normalized by dividing it by the sum of squared distances $d_{ij}(X)$. (Borg & Groenen 2005, p. 42):

$$\sigma_1^2 = \sigma_1^2\left(X\right) = \frac{\sigma_r(X)}{\sum d_{ij}^2(X)} = \frac{\sum\left[f\left(p_{ij}\right) - d_{ij}(X)\right]^2}{\sum d_{ij}^2(X)} \tag{3.6}$$

The square root of equation 3.6 is then what is known as *Stress-1*(equation 3.7), and reported as an indicator of the goodness of fit of the MDS configuration. This

measure is more commonly used than σ_1^2 for the sake of ease of comparison, as σ_1^2 can reach very low values.

$$Stress - 1 = \sigma_1 = \sqrt{\frac{\sum [f\,(p_{ij}) - d_{ij}(X)]^2}{\sum d_{ij}^2(X)}} \tag{3.7}$$

Stress-1 is diminished by optimizing X in a dimensionality m. A perfect stress of 0 can be obtained for any ordinal matrix representation if it is mapped in $m = n - 2$ dimensions. This, however, is counter-productive: The ideal number of dimensions is a compromise between distorting the data structure by over-compression of the solution space, and plotting the data in too many dimensions, resulting in over-fitting of the noise components. As a rule of thumb, any stress below .20 is acceptable (representing a poor fit); values below $.05$ are considered a good fit, and stress values of $.025$ and below an excellent fit. (Borg & Groenen 2005, p. 42-47) Stress is not an entirely rigid measure, though. "[T]he degree to which an MDS solution can be brought into a meaningful and replicable correspondence with prior knowledge or with theory about the scaled object" is really, what is purposeful.(Borg & Groenen 2005, p. 55)

So stress depends on a number of parameters: More data points and more error in the data increase stress, while a higher dimensionality and a higher number of ties and more missing values decreases stress.

3.3 The model

The model specifies the manner of representation of the data in the solution. All distance representations rest on three axioms: (Borg & Groenen 2005, p. 33f.)

- **Nonnegativity:** If $i = j$, then $d_{ij} = 0$, or else, if $i \neq j$, then $d_{ij} > 0$:

$$d_{ii} = d_{jj} = 0 \leq d_{ij} \tag{3.8}$$

- **Symmetry:** The distance between points i and j, is equal to the distance between j and i.

$$d_{ij} = d_{ji} \tag{3.9}$$

- **Triangle inequality:** The direct distance d_{ij} between points i and j can never be greater than the distance between i and j via a point k.

$$d_{ij} \leq d_{ik} + d_{kj} \tag{3.10}$$

3.3.1 The Euclidean distance model

Distances are computed with the help of the *Pythagorean theorem* $a^2 + b^2 = c^2$ Thus, a Euclidean distance d_{ij} between points i and j in a two-dimensional graph is computed as

$$d_{ij}(X) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2} \tag{3.11}$$

returning the length of the hypotenuse of the right triangle, i.e. distance d_{ij}, defined as "the sum of the intradimensional differences $x_{ia} - x_{ja}$." (Borg & Groenen 2005, p. 39)

Equation 3.11 above can also be written as

$$d_{ij}(X) = \left[\sum_{a=1}^{2} (x_{ia} - x_{ja})^2 \right]^{1/2} \tag{3.12}$$

which can be applied to an infinite number of m dimensions as

$$d_{ij}(X) = \left[\sum_{a=1}^{m} (x_{ia} - x_{ja})^2 \right]^{1/2} \tag{3.13}$$

(Borg & Groenen 2005, p. 39)

Weighted Euclidean distance measure

One of the main problems in data analysis is the *Problem of Aggregation*, i.e. how to "appropriately represent the variation in a set of individuals' data". (Coxon

2009, p. 5) Answers to this issue span between the two antonymous approaches which either assume that individuals are unique to such an extent that comparisons between their data are impossible, or, that all individuals' data is representations of one underlying structure, distorted by random variation (i.e. error).

INDSCAL, Weighted Euclidean MDS, is based on the assumption that individuals and groups may have some distinct perspectives, while they still share some common features with others. The Problem of Aggregation is tackled by presuming a *Group Space* X_{ij}, which consists of a fixed set of dimensions, and a *Subject Space*, in which the dimensions constituting the Group Space are appointed a weight between 0 and 1. ($[0...w_{ia}...1]$). The weights can be interpreted as *importance* or *salience* of a dimension, and according to the pattern of these (i.e. their relative importance), a subject can be described in the form of their individual Subject Space.

3.3.2 The PROXSCAL algorithm

Simple Euclidean distance MDS and Weighted Euclidean distance are implemented in PASW (former SPSS) within the PROXSCAL function. PROXSCAL "performs multidimensional scaling of proximity data to find a least- squares representation of the objects in a low-dimensional space.[...] A majorization algorithm guarantees monotone convergence for optionally transformed, metric and non metric data under a variety of models and constraints." (Meulman et al. 2001, p. 296) It is preferable to its alternative, the ALSCAL algorithm, as the latter is prone to distortions and exaggerated error. (Coxon 2009)

3.4 Interpreting MDS output

MDS can be used to explore or to explain data. The former is used to create a graphical representation of the data, whereas the latter is employed to show the structures underlying the data. (Coxon 2003) Under this premise, MDS can be used as a psychological model that transforms judgments of similarity into metric distances.

> The most common approach is to hypothesize that a person, when asked about the dissimilarity of pairs of objects, acts as *if* he or she computes a distance in his or her "psychological space" of these objects. (Borg & Groenen 2005, p. 11)

The underlying structure of the data becomes apparent by analyzing the visual output of an MDS model by interpreting its regions, which can be:

- axes
- clusters
- regions
- manifolds
- a surface in some n-dimensional space
- dimensions

(Borg & Groenen 2005, p. 4f.)

Interpretation of a dimension is achieved by the identification of points distant from one another, of which some qualities are already known; based on the prior knowledge of these characteristics, a substantive criterion that could have induced experimental participants to distinguish between these objects, i.e a criterion that could have led them to place the stimuli at opposite ends of a dimensions, is determined. (Borg & Groenen 2005, p. 11)

3.4.1 Manipulating the graph

All *Similarity transformations*, which are transformations that preserve the distances between points of an MDS representation are permitted for MDS representations on all levels of measurement. These admissible transformations consist of rigid motions, also called *isometries*, like rotation, flipping, and *translation*, the "displacement of an entire configuration relative to a fixed point". (Borg & Groenen 2005, p. 23), of the object space and *Diltations*, i.e. resizing of the entire configuration.

However, Weighted Euclidean distance scaling poses a special case: the dimensions are *rotationally unique*. Since the orientation of the configuration is determined by the weights appointed to dimensions by experimental participants, rotation means the automatic loss of that information. (Borg & Groenen 2005, p. 492)

3.5 MDS for TTS evaluations

Meulman et al. (2001) refers to MDS as "most appropriate when the goal of your analysis is to find the structure in a set of distance measures between objects or

cases".(p.13) This is indeed what we aspire in the evaluation of speech quality: we want to understand the structure of the components that contribute to what we then perceive as the degree of naturalness of synthesized speech.

It has not been long ago that MDS has been discovered as a suitable tool for understanding "what acoustic cues listeners attend to by default when asked to evaluate synthetic speech". (Mayo et al. 2005, p. 1)

3.5.1 Weighted MDS for evaluation of speech quality

Hall (2001) conducted a row of experiments to test Weighted MDS as a stable method for evaluation of synthesized speech. The potential to determine acoustic cues in perception, the correspondence between MOS of stimuli and their location in the solution space, and a measure for fitting new stimuli into the object space were tested.

One sentence read by a male speaker was synthesized by 10 different codecs, and one sentence read by a female speaker was processed by 7 different codecs. The two types of stimuli thus generated were analyzed separately. The 16 participants in the listening experiment, aged 26 to 67, all had previously been exposed to coded speech. From all coded sentences read by one speaker, stimulus triples were created. Each stimulus triple was then presented by means of a digital signal processing platform. Each triple was played in a loop. By pressing designated (physical) buttons in a response box, listeners indicated which two versions sounded most different from one another.

The responses of each individual participant were recorded, and weighted MDS was performed with the statistical programme SPSS in three dimensions. A methodological problem of the Hall (2001)'s stimulus presentation in triades is voiced by Coxon et al. (1982), as

> subjects often find it an extremely wearisome task, and tend to 'lock in' on a single construct. (p.15)

This bears the question whether consistent judgments of a listener result from an individual's perceptual bias or from fatigue caused by a monotonous task.

The graphical representation of the male and the female object space were similar, but not identical. The reason suggested is that some codecs affect different

voices in different manners. MOS that had been collected in the course of a large-scale evaluation of TTS were found to be highly correlated with dimension 1, and to a lesser extent dimension 2 and 3 of both speakers' object spaces. Based on these correlations Hall (2001) compares the predicted MOS to the actual MOS and finds only a small error in predicted scores. From that he concludes that MDS based on similarity-difference judgments is a useful tool for determining quality of speech. Even though subjects were never asked to make quality judgments, MDS analysis structured the stimulus space in a way that was comparable to results obtained in explicit quality judgments.

Hall (2001) also found that variance accounted for in a listener's answers was negatively correlated to the time it took them to make their judgments. Also, it took participants longer to make their decision in cases when the quality of three stimuli was very similar.

In a second experiment 8 stimuli along each dimension from the male speaker's object space were chosen and presented in the same order as plotted along the axis of each dimension. Listeners were then asked to indicate which attribute changed along that axis, and whether it improved or deteriorated. Listeners were reported to struggle with this task, and attributes named varied greatly. Dimensions were identified as naturalness, noise, and low-frequency content. It was then tested whether objective measures would support these classifications: average spectra of stimuli classified as natural deviated less from the grand average spectrum than those classed unnatural. Stimuli classified as noisy created spectra whose high frequency range was above the grand average, whereas those classified as not noisy had high frequency ranges below the grand average. As for the dimension of low-frequency content, stimuli put at the extremes of the dimension had low frequency contents below and above the grand average of at low frequencies, respectively. Analysis of the female speaker's object space following the same procedure yielded comparable results. From that, Hall (2001) concludes that graphical MDS representation is a good fit to perceptual as well as physical correlates.

One issue that somewhat could jinx the external validity of these findings is the choice of test persons in the experiments: Hall does not specify how big the experimental participants' expertize in the area of TTS really is. If they had not only been exposed to synthesized speech prior to participation in the experiment, but if are experts of TTS, the fact that their perception occurred along discrete dimensions may be rather due to the fact that they have schooled ears and also

know along which dimensions synthesized speech tends to vary. To put my concern simply: if a listener knows the variables that are likely to be involved in their judgments, they are likely to perceive a signal in terms of these variables. Thus these variables are more likely to be found in an expert listener's results than in naive listeners. The labels the listeners gave to the dimensions they had to identify raise concern that they are highly specialized in the field: "high frequency components", "brightness, emphasis on mid and high frequencies", "nonlinear distortion bandwidth", "low frequency content", or "spectral richness"(Hall 2001, p. 2181) is not the language of naive listeners.

3.5.2 MDS for evaluation of speech quality

Mayo et al. (2005) conducted a pilot study in which they tested the suitability of MDS for determining these "acoustic cues listeners attend to" in TTS evaluation. In other words: they conducted a pilot study to determine whether MDS was a valid tool for the evaluation of the quality of synthesized speech, which would yield results beyond mere ranking of the systems, but also allowed insights into the acoustic features that determine listeners' judgments of naturalness.

8 sentences from the TIMIT database (Garofolo et al. 1988) were chosen to be synthesized and subsequently used as stimuli in a listening experiment with 8 participants. They were 27 to 35 year-old native speakers of English, who had previously been exposed to synthesized speech. Each sentence was paired up with every other sentence of the synthesized stimuli. The listeners were presented with stimulus pairs and had to indicate whether the degree of naturalness of the stimuli was similar or different, while ignoring all meaning of the sentence. Naturalness was defined as "how much like 'real speech' the utterance sounded". (p. 2) Each pair consisting of a stimulus A and a stimulus B was presented 6 times - 3 times in the order AB and 3 times in the order BA. A new stimulus was presented 2000 msec after a response to the previous stimulus had been given.

From the responses of the listening test a distance matrix was compiled, in which each cell contained the number of times a stimulus-pair had been labeled as *different*. MDS was performed with SPSS. Identification of dimensions was then attempted by visual analysis of the graphical MDS representation and auditory analysis of the stimuli. The main difference to Hall (2001)'s analysis is that not Weighted Euclidean Distance Scaling, but Identity Euclidean Distance Scaling

was performed, i.e. the values summed over subjects were used, washing out the effect of the weights each individual appoints to the dimensions.

The authors thus identified three clusters in three-dimensional space, whose stimuli were characterized by fairly natural sound, extreme prosody errors, and errors on the unit level, respectively. Also, the authors identified a dimension along which a degradation of overall naturalness was observable.

The authors thus concluded that MDS was a useful means for the overall speech quality of synthesized speech, as well as for exploring acoustic cues most salient in the perception of the synthetic stimuli.

Again, as is the case in Hall (2001), it is unclear how expert the listeners were in the field of TTS. It is questionable how representative the test subjects are of the average listeners.

3.5.3 MDS for large-scale evaluation of speech quality

The methodology employed by Podsiadlo (2007) is largely the same as that of Mayo et al. (2005), but the data the analysis is based on is far more extensive: Podsiadlo (2007) analyzed a distance matrix generated from similarity-difference judgments of the evaluation of the Blizzard Challenge 2007. 16 systems were evaluated by 306 listeners from four distinct backgrounds: speech experts, US undergraduate students, UK undergraduate students, and "'real' people" (Fraser & King 2007). 2-dimensional analysis was performed in SPSS. The two dimensions were defined as good/bad joins, and non-/robotic-sounding speech. To test these labels given to the dimensions, a listening experiment was conducted. The 20 participants were all native speakers of English, aged 17 to 63 from varying educational backgrounds. Prior to testing they were trained as to what joins were and how to distinguish different degrees of robotic sound. In two experimental runs, listeners were asked to rate the systems they heard for the goodness of their joins and the degree in which they sounded robotic, respectively, on a scale from 1 to 7. The ranks attributed to the systems are similar to those resulting from MDS analysis and ordering of stimuli along the identified dimensions.

This replication of Mayo et al. (2005)'s study on a larger scale yields comparable results. This more extensive study is the first one to include listeners from different vocational backgrounds and thus the first study to provide conclusive evidence supporting the claim that MDS is a valid means for the evaluation of

speech quality of synthesized speech. It indicates that naive listeners perceive synthesized speech along of dimensions they do not need to have the slightest declarative knowledge of.

MDS is a powerful tool, whose output is also very readable and accessible when a larger number of objects is represented.

What is still problematic, though, is that claims about perceptual dimensions were made, ignoring the weights individual listeners appoint to these. The open question really is whether aggregated data is representative of *any* listener at all, or whether it creates an artifact.

3.6 Summarizing the appeal of MDS for TTS evaluation

MDS is a fairly robust measure, as it is not susceptible to non-systematic missing data, not confined to a certain level of measure, and distribution-free. The graphical output is easily interpretable, and allows to explain the underlying structure in data. (Coxon 2003) Not only does MDS generate results that are similar to those determined in MOS ratings, but it allows further exploration of dimensions to be made. This is agreed on by all authors of articles introduced above, even though the methods and experiments they conducted to arrive there differ somewhat.

CHAPTER 4

Experiment: Research questions and Design

How should you walk in that space and know
Nothing of the madness of space,
Nothing of its jocular procreations?
Wallace Stevens, *The Man with the Blue Guitar*

As has been shown in the previous chapter, MDS is fairly new in the field of TTS evaluation. To close some of the gaps in our knowledge, 5 research questions were formulated. They will be addressed and attempted to answer with data gathered in a listening experiment, conducted by the author.

4.1 Comparison of NSs' and NNSs' judgments

It will be tested whether NNSs are suitable experimental participants for naturalness listening tests: In previous research it has been found that the perception of NSs of a language differs qualitatively from that of non-native speakers (NNSs), as their mental representations of the target language are not the same. This will surface, particularly when noise is introduced into the speech signal: even proficient NNSs can be expected to show a strong decrease in perceptual performance. (Axmear et al. 2005) This becomes particularly apparent in intelligibility tasks. Bennett (2005) reports that in the course of the evaluation of TTS systems in the *Blizzard Challenge 2005*, NNSs were observed to encounter considerable difficulties in Modified Rhyme Tasks and the Semantically Unpredictable Sentences task. A large extent of NNSs' test results had to be excluded from the evaluation procedure, as participants either did not give answers to all the questions within a task, or gave up on a task altogether. One suggestion is that NNSs were encountering difficulties with words they were not familiar with, which posed

spelling problems as well as it resulted in excess of their memory capacity. The specific nature of these problems NNSs encounter has not been further investigated. (Bennett 2005) The current study tests whether these problems are limited to comprehension, or whether a listener's native language influences their judgment of naturalness of stimuli as well, given that they speak the target language of the TTS system fluently.

4.2 Comparison of results from MOS and similarity-difference ratings

Furthermore, it will be investigates to what extent the results gained in MOS listening experiments compare to those from similarity-difference judgments: As it been established above, MOS are a valuable measure for establishing ranks of overall performance. These ranks will be compared to ranks computed from similarity-difference judgments.

4.3 Relation of individuals' judgments to judgments of a population

Mayo et al. (2005) points out that listeners attention to different dimensions in the acoustic signal varies. To what extent this is done, and how the weighting of dimensions across a sample of experimental participants is distributed, will be investigated in the analysis of subject spaces in the weighted Euclidean MDS. It is a vital question, how representative the individual's judgment is of that of a population. If there is no big digression of a listener from the data generated by and averaged over several listeners, individual weighted scaling is obsolete; furthermore, it would let us hope that the amount of testing that is needed to evaluate TTS systems could possibly drastically be limited.

4.4 Comparison of the outputs generated by direct and by aggregated data

The amount of test stimuli in the Blizzard challenge 2008 is vast, so each experimental participant only gets to listen to comparatively tiny subsets of the test data. Averaging then occurs across subjects and across stimuli for each system. In a listening experiment a small subset of these stimuli will be presented, to

generate full matrices for all participants. The output of various configurations of simple Euclidean distance scaling and weighted Euclidean distance scaling on the subset as well as to averaged data from the large scale evaluation will be compared. If we fail to deliver at least a sufficiently large intersection between MDS on averaged data and weighted MDS, this seriously questions current MDS testing methodology.Finally, it will be investigated as to how well the experimental results gathered from this subset generalize to a larger set of stimuli and test participants. The values of the two sentences for each system will then be averaged, and it will be compared, whether these are comparable to results gained in the more elaborate Blizzard challenge listening test.

The research questions, as they are laid out above can be described along three lines:

- Characterizing the typical listener
- Determining the adequacy of MDS (functions) for evaluating the speech quality of TTS systems
- Determining to the relationship between MDS and the established measure of MOS
- Determining to what degree a small subset of stimuli is representative of the speech quality of a TTS system.

These questions are at the very basis for any research in minimizing the amount of testing needed for MDS analysis of TSS evaluation experiments. As a first basic step First, we must determine what an average listener is, so that we can further scrutinize how they are related to a bigger pool of listeners. The average listener is approached by means of exclusion: every listener is assumed as a valid representative, unless they have shown to have a specific feature that classifies them as otherwise. Podsiadlo (2007) reported that judgments of listeners did not vary significantly between different age groups or varieties of English, so these are not considered as identifiers of atypical listeners. If the factor of native-language competence is found to significantly affect listeners' perception, non-native speakers will have to be excluded from all further general evaluations of TTS generated speech quality. This, however, may bring about a conflict of interests: as TTS systems are employed in an increasing range of contexts, such as for characters' voices in video games, travel information, and booking services for (international) travel (Taylor 2009), they may very well be used by non-native

listeners. The scope of validity of evaluation experiments may be limited even further, depending on the target user group of the system.

Next, it must be established whether the measures gained in MDS correlate with those of test methods which have been tried and tested. Unless this is the case, the suitability of MDS is seriously questioned, and all further considerations are pointless.

Once the average listener is characterized, it will be investigated how representative the judgment of a single listener is of that of the average generated from the judgments of a larger group of listeners. If indeed variation is small, large subject pools may soon be obsolete in evaluation. Finally, the degree to which one stimulus or a low number of stimuli from one system is representative of its overall performance is investigated. As the quality of one system's output can be expected to fluctuate significantly, one randomly chosen stimulus should not be sufficient. However, maybe an average between a good and a bad sentence generated by the same system may approach a fairer representation of performance. Again, if this is the case, further research could optimize the choice of test stimuli, and hence, again, limit the amount of testing that is required for TTS evaluation.

To address these questions, an experiment consisting of two parts has been devised. Part 1 gathers data for MDS analysis, by asking participants to judge Part 1 gathers subjects' judgments of similarity and difference between the naturalness of samples, whereas part 2 collects mean opinion scores (MOS) of perceived naturalness of the samples.

4.5 Participants

Altogether 40 participants from different vocational backgrounds were tested, 30 of which were NSs of some variety of English, while the remaining 10 were NNSs of English (from now on for reasons of simplicity referred to as NNSs). The NNSs were fluent in English, and their first languages were from different language families. It will have to be investigated, whether, and if so, to what degree there is significant variance within as well as between groups of native speakers and non-native speakers. This will also determine, whether it is necessary to exclude a certain group of speakers from further analysis, as they must be considered as atypical speakers. The pool of participants was self-selecting: Participants were

label	type	sentence	syllables	duration
T1	natural	For good measure, he offered an unreserved apology.	15	2.9s
T2	synthesized	Billy could help Saxon little in her trouble.	12	2.2s
T3	synthesized	UCA based air traffic controllers are also unsettled.	15	3.4s
T4	synthesized	We are pulling on in the morning to circle city.	14	2.2s
T5	synthesized	I believe the two years suspension are harsh.	11	2.4s
B1	natural	Power cuts affect refrigerated medicines and food stuffs.	15	3.4s
B2	synthesized	But they can live in a pigsty.	8	1.38s
B3	synthesized	He was puzzled by the slowness of its progress.	12	2.2s
B4	synthesized	Thus he waited, keeping perfectly quiet.	11	2.4s
B5	synthesized	The bloodshed was not confined to Copenhagen.	12	2.5s

Table 4.1: Stimulus sentences

chosen on a first-come, first-serve basis in their response to an advertisement. They were paid 7 to take the experiment, which took none of the participants longer than 40 minutes to complete.

4.6 Stimuli

Hall (2001) suggests 8 or 10 samples of synthesized speech "that span the perceptual space under question' (p. 2168) are chosen as stimuli. The 10 stimuli used in this experiment are a subset taken from five participants in the Blizzard challenge 2008, test set A. Since the Blizzard Challenge is a large-scale evaluation of TTS systems, it is not feasible that one experimental participant is presented with all stimuli that are used for evaluation, as the quantity is sheer overwhelming. So this experiment was devised to have full, square distance matrices of all systems in the smaller scale experiment compiled for each participant. This allows a controlled comparison of the weights individuals appoint to dimensions, as well as a direct comparison of the output generated by Euclidean distance measure and Weighted Euclidean distance measure MDS.

The four systems chosen for comparisons were selected on the basis of four representative sentences that had the lowest difference scores in direct comparison with naturally recorded speech in a subset of the data collected in the Blizzard challenge. For each system, another sentence that is perceptually more distant from natural recorded speech was added, as well as two sentences of natural recorded speech.[1] Natural speech recordings were included to "anchor the scale". (Taylor 2009, p. 537) The length of the sentences ranges from 1.38 to 3.4 seconds, and from 8 to 15 syllables.

[1]The stimuli used in the experiment can be accessed at http//:homepages.inf.ed.ac.uk/s0674876/listening_test_july_2009_wavfiles/

Taylor (2009) report that instead of rating naturalness in MOS tests, listeners tend to indicate how much they like a particular system. Naturalness and likability often do go hand in hand, but in situations when they are presented with an unpleasant natural sounding voice, and a less natural, more pleasant sounding voice, the latter tends to be given a higher score. (p. 536) This effect is eliminated in this experiment, since all voices are made from the same original recordings. Ideally, the natural recordings should be given a perfect rating of 10.

4.7 Procedure

The experiment is conducted in a computer lab. Instructions, as well as stimuli are represented on the 20 inch screen of an imac computer as a web page within a Firefox browser window on full screen mode. Answers are given by clicking the respective radio-button on the screen, using an optical mouse. The subjects listen to the stimuli with closed-back Senheiser headphones and at a volume level can they adjust themselves.

4.7.1 Part 1

Each stimulus was paired up with every one of the other stimuli, so that paired comparisons between all stimuli were made, in both direction. This was done, because the order of presentation within a stimulus pair could also affect subjects' perception of similarity and difference between sentences. The stimulus pairs were presented in random order. Participants had to decide whether both items of the pair were equal or different in their degree of naturalness of their sound. As in Mayo et al. (2005)'s pilot study, listeners

> were not instructed to listen to any one acoustic characteristic of the stimuli, or to any specific psychoacoustic construct (e.g., listening effort, pleasantness, pronunciation etc) such as have been used in previous evaluation studies e.g., [Sluijter et al. (1998)]. The task was simply to make a simple binary decision about the degree of similarity in naturalness of each pair of stimuli.

4.7.2 Part2

Part 2 was devised to rank the systems according to their naturalness. Ranking stimuli can be done by asking experimental participants to listen to stimuli and

put them into order according to some quality. This task is usually perceived as very hard, and listeners tend to be more effective when scoring a stimulus, as in a MOS task. These scores can then be ranked. The listeners' scores are then recorded as conditional similarity data, which means that values cannot be compared directly between subjects. (Coxon et al. 1982, p. 14) Each stimulus was presented three times in random order. Participants had to rate on a scale from 1 to 10 (1 being the lowest, and 10 the highest), how natural a sentence sounded. Generally, in MOS tasks, measures between 1 and 5 are used. (Holmes 2001, Hall 2001, Jurafsky & Martin 2008) However, for this experiment, consciously a larger range was chosen so that the ratings would be a bit more dispersed, in the hope that this will generate bigger gaps between the systems' ratings and that distinct ranks could be clearly established. An ideal-case scenario is that the two natural stimuli, B1 and T1 receive perfect scores of 10, each, as they *are* natural voices, and therefor should also be perceived as natural sounding.

CHAPTER 5

Experiment: Analysis and Discussion

We shape our tools and thereafter our tools shape us.
Marshall McLuhan, *Understanding Media*

The data of part 1's 40 participants, resulting in 400 cases for 10 objects, which are 4000 edges, out of which 29 were missing, was put into a full distance matrix. The proximities are stacked in 10x10 matrices across columns. *similar* judgments were coded as 0, *different* judgments as 1. The experimental results were analyzed with PASW Statistics 17.0 (formerly SPSS Statistics). MDS graphs were generated with the PROXSCAL function, which includes the *Identity Euclidean* function, and the *Weighted Euclidean Distance* function.

The data in part 1 is defined by dichotomization. So given three stimuli *i*, *j*, and *k*, it is true that if *i* and *j* are similar, and *k* is different from *i* and *j*, then $d_{ij} < d_{ik}$. However, some of the axioms of distance representation, as explained in chapter 3, are not obeyed to the letter:

The axiom of symmetry (cf. Formula 3.9) assumes that both triangles of a matrix of proximites (i.e. the upper and the lower triangle of the matrix) are symmetrical, i.e. that the order of presentation of stimuli is irrelevant.

In Figure 5 it is visible that the lower triangle tests stimulus pairs in which the lower-numbered stimulus is the initial item of the pair, while the upper triangle tests pairs constructed the other way round: In clearly symmetric data, this should have no effect at all. We suspect, however, that our data will not be symmetrical, and that order of presentation will have an effect on similarity difference judgments. A comparison of MDS representations generated from the

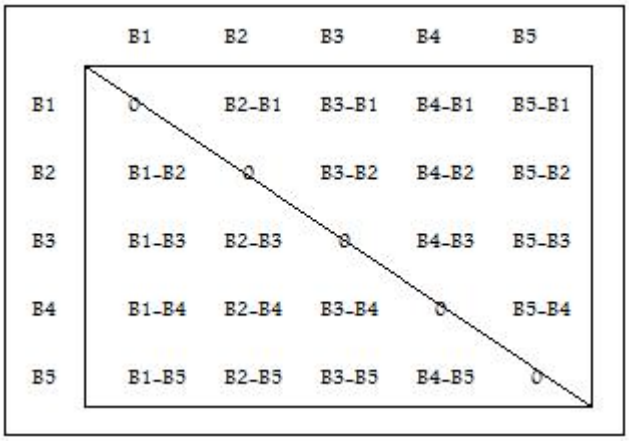

Figure 5.1: Upper triangle vs. lower triangle of the matrix as stimulus input: tests stimulus comparisons in different directions

Test Statistics[a]

		B1	B2	B3	B4	B5	T1	T2	T3	T4	T5
Most Extreme Differences	Absolute	.200	.300	.300	.400	.300	.200	.100	.300	.300	.200
	Positive	.200	.300	.300	.400	.000	.100	.100	.100	.300	.200
	Negative	-.100	-.100	-.200	-.100	-.300	-.200	-.100	-.300	.000	-.200
Kolmogorov-Smirnov Z		.447	.671	.671	.894	.671	.447	.224	.671	.671	.447
Asymp. Sig. (2-tailed)		.988	.759	.759	.400	.759	.988	1.000	.759	.759	.988

a. Grouping Variable: TRI

Table 5.1: Kolmogorov-Smirnov Z test determining whether there is a significant difference between distances of points as defined by MDS of the upper and the lower triangle of the input matrix

two different triangle matrices was conducted. For that effect, all participants in whose data there were one or more cells empty, were eliminated. This was done to make sure that an equal number of judgments generated both matrices.

For each of the triangles, Weighted Euclidean distance scaling was performed in two dimensions on an ordinal data level, untieing ties. The distributions of distances generated from each input triangle are normal, with the exception of B4 in the upper triangle, for which a Kolmogorov-Smirnov test of Normality proved significant a $\alpha < 0.05$. Cells of the upper and the lower triangles are positively correlated at a significance level of $\alpha < 0.05$.

When comparing the sums of distances of the upper and the lower triangle, it is obvious that there are differences between judgments (cf. Table A)

A Kolmogorov-Smirnov Z test was conducted that indicated that none of the differences for the variables was significant. (cf. Table 5)

rank	LOWER	UPPER
1	T1	T1
2	B1	B1
3	B4	B3
4	T5	B4
5	T4	T5
6	B3	B5
7	T2	T3
8	B2	T2
9	B5	T4
10	T3	B2

Table 5.2: Ranks of stimuli as appointed to distances between stimuli and T1, generated from lower and upper input matrix of Weighted Euclidean MDS

However, when stimuli are ranked according to their distance from T1, resulting ranks differ: (At this point, a word of caution: whenever we deal with ranks, it has to be borne in mind that they are not an exact measure, as they are retrieved from distances distorted by noise. (cf. (Borg & Groenen 2005, p. 53ff)) We consider rank computation as a valuable means for getting insights into overall tendencies of perceived quality of stimuli.)

The biggest difference between the two ways of ranking the stimuli is that between the ranks given to T4.

Acknowledging the fact that the present matrix is asymmetrical, we will for further analysis always use the full matrix. The other axioms of nonnegativity and triangle inequality can be flouted, if ordinal measures are used for MDS representation. Ordinal MDS will in all further analysis provide us with robust measures, as well as keep stress values acceptable also in low-dimensional representations. Fewer dimensions will render the visual analysis of MDS graphical output considerably easier.

MDS representations were generated with the PROXSCAL function, which includes the *Identity Euclidean* function, and the *Weighted Euclidean Distance* function. PROXSCAL can deal with missing values, and thus participants with missing cells are included into analysis again.

5.1 The typical listener

"As a methodological principle, the inspection of individual differences should always precede aggregation". (Coxon et al. 1982, p. 15) Systematic differences of individuals can otherwise be lost, and if different subgroups' biases balance one another out, the final representation can be an artifact that is not representative of any of its groups. Some schools assume the variation in individual participants is noise which is eliminated by aggregation. (Coxon et al. 1982) This assumption is supported by Simple Euclidean MDS, but not Weighted Euclidean MDS. How these two approaches compare will be investigated below. However, to even reach that point of discussion, we first need to determine whether there is a systematic perceptual difference between subgroups of participants.

For that reason, the first step needed in this analysis is a comparison of the two listener groups of the experiment: NSs and NNSs. The initial analysis is a Kolmogorov-Smirnov Z test, investigating whether the difference judgments of NSs and NNSs are from one population, and whether their MOS are from one population. Kolmogorov-Smirnov Z tests will be our test of choice in most of our analysis of the effect of native language: it is non-parametric and thus not sensitive to (the lack of) normal distributions and homogeneity of variance. Also, it is preferable to the more common Mann-Whitney test, because of the low sample size of only 10 NNSs. (Field 2005, p. 529)

The MOS of NSs and NNSs are from one population (cf. Table A), as are their judgments in part 1, with the exception of similarity-difference judgments concerning stimulus T3, for which differences are significant at a level of $\alpha < 0.05$. (cf. Table A). To test what effect these differences have on MDS representations, we conducted our first MDS analysis *including* NNSs' data.

Weighted Euclidean Distance MDS was performed on the ordinal level, untieing ties, applying transformations to each point individually. Stress 1 is 0.17, which is an acceptable fit. Stress decomposition was performed. Normalized Raw Stress values for participants ranged from 0.0048 to 0.1015, and the distribution of stress values is normal. However, it is interesting to note that the higher stress values were occupied by NNSs; 9/10 NNSs' stress values were above average. This suggests that NNSs dilute our results by introducing higher amounts of variance in the data than NSs. This assumption was tested with a Kolmogorov-Smirnov Z test. The hypothesis that the stresses introduced by NSs and NNSs are not from one population is significant at $\alpha < 0.01$. (cf.Table A)

Already the first step of analysis indicated that the judgments of T3 should be categorically different for NSs and NNSs, and variance in the NNSs' judgments increases the potential for error (i.e. stress). How much variance influences stress can be exemplified at the example of B4. Above, we have found that the upper and the lower triangle for B4 were significantly different, i.e. the position of B4 in a stimulus pair had a significant effect on how the stimulus was perceived. This results in great variance in the data, which in turn results in higher stress values. In our MDS representation relying on the full matrix as input, we analyzed normalized raw stress values, which ranged from 0.0223 to 0.0336: stimulus B4 had the highest stress value.

Even if it was not for the divergence in the case of T3 NNSs introduce, larger amounts of variance in the judgments, which in turn necessitates a larger number of listeners in order to determine the underlying pattern representative of the population. Increasing the numbers of participants required in an experiment is not desirable at all, as every additional participant creates additional costs. For this reason, we argue in favour of the exclusion of native speakers for all further analysis.

It is vital, though, to note that a Kolmogorov-Smirnov Z test scrutinizing whether the weights NSs and NNSs appointed to different dimensions were from one distribution found no significant difference between the two groups. This result suggests that the judgments obtained by NSs are representative of NNSs, and therefore no distinct large-scale evaluation for NNSs is needed, once one for NSs is in place.

5.2 Interpreting the MDS space

Having excluded the data of NNSs, Weighted Euclidean MDS was performed in two dimensions on an ordinal level, untieing ties, applying transformations to each point individually. Stress-1 is 0.14, and Dispersion accounted for (D.A.F.) is 0.98, which is a reasonable fit. So for now we will limit ourselves to two dimensions in favour of ease of interpretation of the graphical representation of the stimulus space. Analysis is done visually and auditively. We attempt to organize the stimuli into clusters according to their auditory features. The two natural recordings, T1 and B1, are clustered together clearly distinct from the other stimuli. Hence we can deduce that experimental participants perceived a clear distance between those and the synthesized stimuli. This supports Holmes (2001)'s

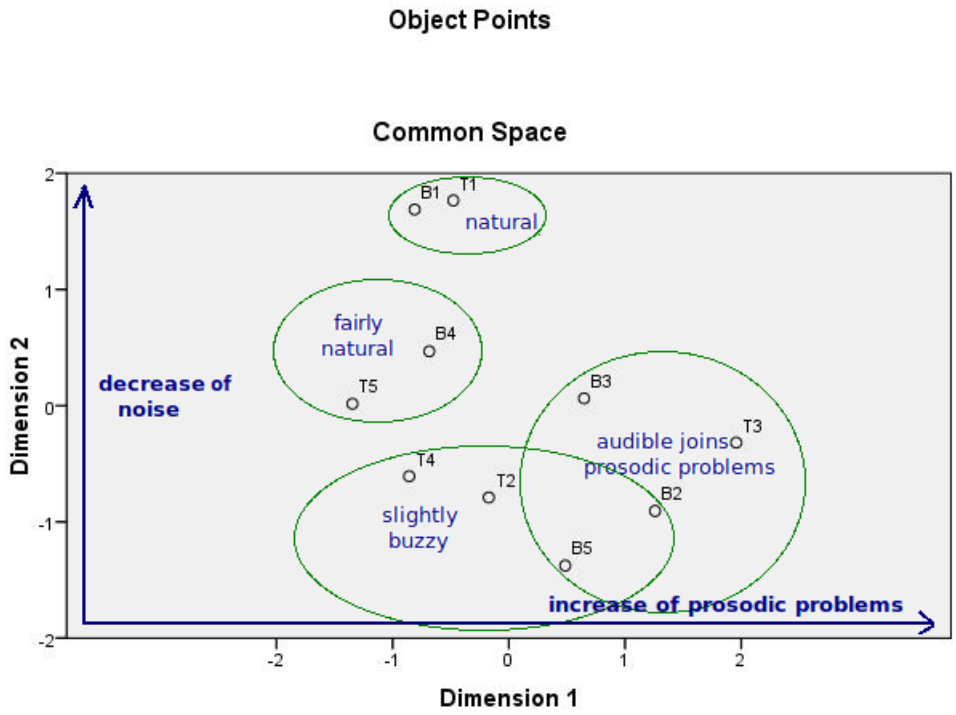

Figure 5.2: 2-dimensional MDS space generated from weighted ordinal MDS of NSs' judgments

claim that "even the best examples of speech from TTS systems are unlikely to be mistaken for natural speech". (p. 107) Also, the two natural recordings are closer together than any of the TTS systems' sentences (or any other two sentences, for that matter). This can be explained by the fact that the quality of natural speech is fairly constant, which cannot be said for TTS systems. Putting it simply, we could say that at its best, a given TTS systems is very good and comes rather close to sounding like natural speech, while at its worst, it can be quite dreadful. This is particularly striking in the samples of B5-T5. B5 is much more distant from B1 and T1 than the system's corresponding sentence T5.

In the higher area of dimension 2 (in further text referred to as d2), there is only one cluster, namely that of the natural stimuli B1 and T1. Then we can make out a cluster further down d2, consisting of T5 and B4, both of which sound fairly natural, but still are perceivably artificial. The remaining 6 stimuli fall into two clusters, in the intersection of which B5 and B2 are located: B5 and B2 are the most unnatural of the test stimuli; they are characterized by a certain buzzy quality about them (left cluster, consisting also of T4 and T2), and bad prosody

(right cluster, consisting also of B3, and T3). Based on these observations, we can appoint qualities that increase/decrease along the axes of the 2-dimensional MDS space: for doing so, we will look at points that lie on a straight line, parallel to one of the axes and try to identify a characteristic of all stimuli that increases/decreases along that line: the first line we draw extends from B1 through B4 to T4. It is observable that as we progress form T4 to B4 and B1, the speech becomes increasingly clear, and rid of noise distorting the signal. To see, whether that applies all throughout the space, this vertical line we constructed is shifted parallelly, and now goes through B3 and B5. We expect to find that B5 is noisier than B3. B3 is indeed is less noisy than B5 insofar as it does not have the same "echoing" quality to it.

Analysis of dimension one (in further text referred to as d1) is made by drawing a line between T4, T2, and B2. While T4 has a somewhat buzzy quality, its prosody is fine. In T2 there is definitely something wrong with prosody, but it is hard to say whether it is mainly a problem of intonation, emphasis, or segment duration, as the three are interlinked tightly. The phrase-final syllable, consisting of a syllabic *l*, sounds somewhat clipped. The very same problem is present in B2, but it is more striking, as the final syllable is an open syllable consisting of a diphthong, which should be long. Thus the problem is more salient. Shifting the line to go through T5, B3, and T3, we again find a prosodically acceptable stimulus, one which has an intonation/emphasis/segement duration problem, and one with a particularly striking duration problem, respectively. T3 is an interesting case: it is located at the edge of the NSs' stimulus space, as the most extreme point along d1. It is fairly natural sounding in general, but it has one grave duration error, which is rendered more salient by its position: the voice-onset time of the *t* in the word *controllers* is so long, it almost seems like a break half-way into the word. A break like this would not occur in naturally spoken English, though, since even when breaks are made within words, they tend to be made between syllables, but never in a syllable's onset, as it is the case in T3. Above, we found that the judgments of this stimulus are significantly different for NSs and NNSs. This suggests that while NSs found this error in duration very disturbing and unnatural, NNSs were more forgiving of that fault: the average in the T3 cell for similarity-difference judgments is 0.65 for NSs, while it is only 0.47 for NNSs. We may hypothesize from this that NNSs essentially display perceptual behaviour comparable to that of NSs, but that errors in duration and stress are not frowned upon to the same extent (possibly also because NNSs are not as confident of their own judgments of durations and stress placement).

Overall naturalness is visible as the distance between T1 or B1 and the respective stimuli.

Having been able to create a stimulus space that is organized along the axes of perceptual dimensions from similarity-difference judgments substantiates claims already made by Hall (2001), Mayo et al. (2005), Podsiadlo (2007): MDS allows to organize synthesized speech stimuli according to their naturalness, on the sole basis of data generated from similarity-difference judgments. Since we succeeded in identifying the perceptual dimensions that define the stimulus space, these dimensions enable us to gain insights into how these factors influence judgments of naturalness. However, the authors used different MDS models to reach that conclusion. So far, we have only employed Weighted MDS, like Hall (2001). We will now investigate whether Simple Euclidean MDS generates output comparable to that of Weighted Euclidean MDS.

For that purpose, I conducted Simple Euclidean MDS on an ordinal level, untieing ties. Stress 1 is 0.15, which is an acceptable fit, and D.A.F is 0.98. The output graph (cf. Figure 5.2) resembles that of Weighted Euclidean MDS. It only needs to be flipped vertically and rotated, to generate a representation like the one we just analyzed above. Now it is possible to mark the dimensions of noisiness of speech signal and quality of prosody as done previously. The only difference between the two representations is that in Simple Euclidean MDS B5 is slightly higher up on d2 than it is in the other graph. From this we can conclude that the tested group was sufficiently homogeneous to generate comparable graph outputs for Simple and Weighted Euclidean MDS. This supports Mayo et al. (2005)'s assumption that the MDS representation they generated actually is representative of an average listener and not just an artifact resulting from the interference of different groups' perceptual patterns.

5.3 MOS

Correlations between dimensions and MOS scores were tested: For the Weighted MDS, there was a significant negative correlation between d1 and MOS, r=-.802, p(two-tailed) < 0.01, and a significant positive correlation between d2 and MOS, r=.847, p(two-tailed) < 0.01. There is no significant correlation between d1 and d2, which indicates that in our analysis of the stimulus space above we have indeed identified two discrete factors that influence listeners' judgments of speech quality.

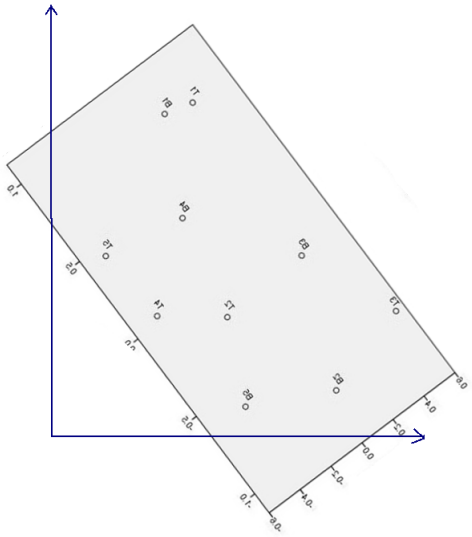

Figure 5.3: Two-dimensional Simple Euclidean MDS, ordinal level, untieing ties, flipped and rotated to align to the axes of Weighted Euclidean MDS representation

The linear regression results for Simple Euclidean MDS allowed fewer insights along the 2 dimensions: d1 and MOS had a significant positive correlation, r=.98, p(two-tailed) < 0.01, and no significant correlation of d2 and MOS. There may be dimensions that fit MOS scores even better, which can be found by rotating the space. When comparing the ranks computed from distances generated by Simple Euclidean to ranks as they are established by other MDS models and their correlation with MDS, it is indeed the one that comes closest to duplicating the ranks generated from MOS scores.

Ranks were computed for two-dimensional MDS representations of Weighted Euclidean distance measures with transformations applied to weights individually as well as with weights transformed simultaneously, MDS representations of Simple Euclidean distance measures, as well as for MOS scores. (cf. Table 5.3) In the MOS task, each participant rated each stimulus three times. The values for these three were averaged to generate one value per sentence per participant.The two natural sentences clearly received the two highest ratings. (cf. Table A) It can thus be said that they are clearly perceived as more natural sounding than the systems. The ranks for the other systems were established by averaging the scores of the two sentences and then they were rank-ordered. (cf. Table A) As T1 received ratings closer to the perfect score than B1, it was adopted as the prototypical natural sounding sentence. The similarity-difference subjects in part 1 were then transformed into ranks for each system by averaging the values of

rank	Weighted Euclidean	wE, simlut. trns	Simple Euclidean	MOS
1	T1	T1	T1	T1
2	B1	B1	B1	B1
3	B4	T5	B4	B4
4	T5	B4	T5	T5
5	B3	T2	B3	T4
6	T4	B5	T4	T2
7	T2	T4	T2	B3
8	B2	B3	T3	B5
9	T3	T3	B5	T3
10	B5	B2	B2	B2

Table 5.3: Ranks of stimuli, computed as their distance from stimulus T1 in two-dimensional Weighted Euclidean MDS and Simple Euclidean MDS, respectively, and ranks from MOS tasks

their distance of their two sentences to stimulus T1, as generated in a weighted Euclidean MDS measure. According to Hall (2001), the MOS, as were collected in part 2 should have correlates in the output of part 1.

Even though ranks only give a rough approximation, and are very vague, they show are that there is a certain consensus across all MDS models as well as MOS scores, and the rough order is very similar. Even though some stimuli vary slightly in their ranks, for each stimulus there is no question whether it is more in the front, the middle, or the back of the field.

To further investigate correlations between positions in the stimulus space and MOS, several linear regression tests were run: the distances computed by different MDS models, as specified above, were the respective independent variables, from which the dependent variable, the MOS and the MOS ranks, were predicted. The best result for both MOS measures was achieved by Simple Euclidean MDS, with transformations applied individually, with an R-square of .911 and .899, respectively, at a significance level of $\alpha < 0.01$. This means that the MDS representation accounts for 91% of variation occurring in the MOS. A reason for the best fit of Simple Euclidean MDS with individual transformations may be the fact that averaging over stimuli occurred here, just as it did for MOS scores. In Weighted Identity MDS, simultaneous transformations accounted for a better fit in the decimal place of per cent of variation accounted for in MOS, and achieved an equal fit in accounting for MOS ranks.

Multiple linear regression was then performed to further investigate the nature of the relation between an MDS representation and MOS. As in two-dimensional Weighted Euclidean MDS d2 has been shown to have a stronger correlation with MOS than d1, d2 was used as first input variable in blockwise entry of variables in linear regression. d2 accounts for more than 70% of variability in MOS scores,

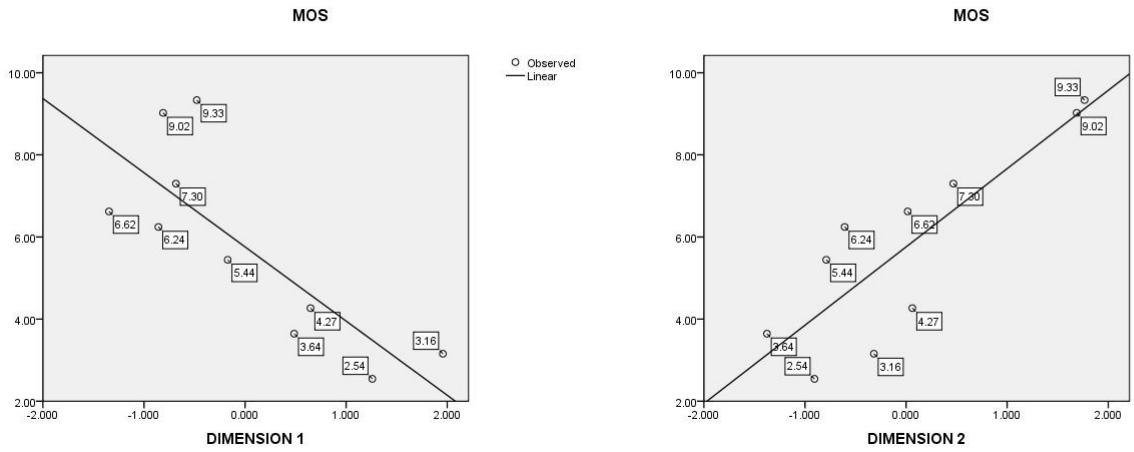

Figure 5.4: MOS scores plotted vs dimension 1 (left) and dimension 2 (right)

and together, d1 and d2 account for 95.4%. The regression was a good fit ($R^2_{adj} = 94\%$), and the overall relationship is significant ($F_{2,7} = 72, p < 0.01$). With other scores held constant, MOS were positively related to dimension 2, increasing by 1.387 for every unit in dimension 2, and negatively related to dimension 1, decreasing by 1.212 for every unit in dimension 1. All effects were significant at $p < 0.01$ Thus we can estimate MOS from our MDS representation as follows:

$$MOS = 5.758 + (1.387 * d2) - (1.212 * d1) \tag{5.1}$$

The rms error between measured and predicted MOS is 0.579 per unit. This score is not at all bad; when comparing it to that achieved by Hall (2001), we must bear in mind that MOS in this experiment ranged from 1 to 10, as opposed to Hall's range of 1 to 5. Consequentially we are faced with bigger variance, which influences rms. This configuration in fact outperforms the regression of 3-d Weighted MDS and MOS ($R^2_{adj} = 93, 2\%, F_{3,6} = 42, 151, p < 0.01,$ rms at 0.619 per unit). This suggests that two-dimensions are a fairly suitable configuration for fitting the external MOS ratings.

5.4 Comparing direct and aggregated data

A subset of similarity-difference judgments collected at the Blizzard challenge was used for MDS analysis. In the evaluation of the Blizzard Challenge, judgments are aggregated across listeners as well as across stimuli of one system. In order to avoid including any further noise in the data, only NSs' judgments were

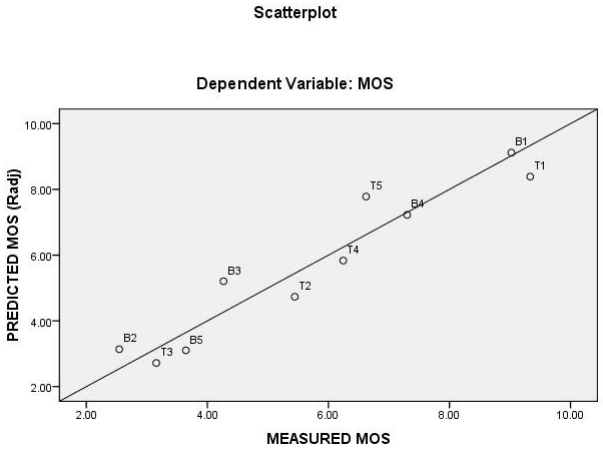

Figure 5.5: MOS predicted from two-dimensional linear regression

evaluated, and the number of listeners within each test group and the number of stimuli each listener judged were kept constant. This meant that a lot of listener judgments were lost, particularly because for a large number of speakers the variable of native language had not been defined. However, it seemed the smaller price to pay. Listeners, whose judgments were included in the analysis, were chosen as follows: during the experiment, each listener was appointed to a test group, which determined the subset of stimuli they were presented with. For the analysis, listeners were sorted within the groups according to their case ids. Then all listeners, who had not completed 21 similar-different judgments were excluded. The numbers of remaining listeners within each group were compared, the lowest of which was 6. The data generated by the first 6 listeners from each group was processed for MDS evaluation, all other data was discarded. As before, *similar* judgments were appointed a value of 0, *different* judgments a value of 1. These values were summed up across participants across groups across stimuli for each system. The resulting 21x21 matrix was then used as input into PASW.

Two-dimensional ordinal Simple Euclidean distance scaling was performed, untieing ties, resulting in a stress-1 of 0.22, and D.A.F of 0.95. The stress level is above what generally is considered acceptable, but for the sake of ease of comparison with our previous MDS representations we will work with this one, anyway, rather than plotting it in more dimensions.

Interestingly, there is a match between the ranks computed for the systems on the basis of distances derived by MDS from the aggregated Blizzard data, and the

ranks computed from distances of Weighted Euclidean MDS on an ordinal level, untieing ties, applying transformations simultaneously, in our smaller experiment. This supports the assumption that aggregated data will indeed provide reliable results that are representative for a larger population. This is supported by the fact that if only one stimulus per system is considered (e.g. all B stimuli for a system), the resulting ranks can be very different. This supports out initial hypothesis that averaging over a good stimulus and a bad stimulus of a system could indeed be a good approach towards approximating the results of larger scale evaluations. Further testing on a larger scale will be needed to investigate how reliable that measure is, but these initial findings here are very promising.

CHAPTER 6

Conclusion

"I want to understand everything," said Miro. "I want to know everything and put it all together to see what it means."
"Excellent project," she said. "It will look very good on your resume."
Orson Scott Card, *Speaker for the Dead*

The main problem of subjective evaluation of speech quality has been defined by the composite nature of speech quality. Listeners know whether a system sounds natural, or in a comparison, which system sounds more natural than another system, while being unable to reliably isolate the factors that brought about their opinion. When employing objective methods, the acoustic differences between systems can be found, while they are not perceptually salient. This problem can be tackled by using MDS.

The findings of the experiment conducted shed some light into the still thick darkness that is MDS in the evaluation of quality of synthesized speech. I succeeded in identifying distinct perceptual dimensions along which the quality of stimuli increased/decreased. By eliminating NNSs from the test group, the data became homogeneous enough to gain comparable results from Weighted Euclidean MDS and Simple Euclidean MDS. This implies that the single listener is representative of the average of a whole group. By averaging the data of a good example and a bad example for each system, an MDS graph that is comparable to that of much larger scale evaluation has been created.

These results suggest that if a group is homogeneous enough - which a group of NSs of English seems to be - configurations relying on data that is aggregated across listeners is an acceptable representation of the single listener. If stimuli are chosen appropriately, the aggregation of data generated by a small number can

be representative of a larger number. This hypothesis must still be tested more extensively, but if it holds, the amount of data required for evaluation can be drastically minimized. Even though complete data matrices are intuitively more exact, an approximation relying on less data may actually be desirable for many reasons:

> A price is paid for data, not only in financial terms but in wear and tear on the organism at source. A method with too high a channel capacity may, through boredom and fatigue, result in a decrease in information transmitted, through stereotype of behavior. Furthermore, the potential variety of messages from the organism may not be great, in which case a more powerful method is inefficient. [...] Ideally a method should be selected which matches the information content in the source but is not such a burden as to generate noise. (Coombs 1964, p.51)

6.1 Outlook

So what implications does that have for evaluation projects like the Blizzard Challenge? If further tests in averaging stimuli of a system prove successful, it should indeed be feasible to reduce the number of participants needed in the MDS part of evaluation. However, in order to do so, the experimental environment will need to be very controlled: to reduce variance of the judgments, participants should all be NSs of English and tested in the same, quiet environment, using the same equipment. Recordings of the time elapsed between initial presentation with a stimulus and the point when a decision is entered could be used as a further measure for checking distribution/dispersion of stimuli (cf. Hall (2001)) as well as of listeners. MOS, which are collected anyway, can also be used to check for individual's biases, and collectively as a reference frame to check whether the resulting MDS representations are plausible. The stimuli chosen as representatives of a system can be derived in a test series prior to the main evaluation: Similarity-difference tests are conducted in the same manner as used in the experiment described in this paper, using a natural stimulus and a few test sentences from a system. MDS is performed and the sentence with the biggest and that with the smallest distance from the natural system is selected. This is done for all systems to be tested, and the thus selected stimuli are then used in large scale evaluation. This part in itself is a fairly expensive again, but the stimuli thus picked will remain representative of a system, and

this part of testing will not have to be repeated, unless changes are made to the system. Hence, in future years, it will be less costly to include more systems from previous years into the Blizzard Challenge for the sake of anchoring. This small step towards a benchmark is a great improvement in subjective evaluation of synthesized speech.

APPENDIX A

TRI	B1	B2	B3	B4	B5	T1	T2	T3	T4	T5
LOWER										
Sum	17.807	19.165	14.078	14.403	17.080	22.088	15.812	21.947	15.548	18.072
Mean	1.781	1.916	1.408	1.440	1.708	2.209	1.581	2.195	1.555	1.807
StdDev	.873	.998	.662	.772	.945	1.176	.862	1.207	.918	1.056
Kurtosis	.51	-.10	1.93	.67	.32	-.18	-.29	-.51	-.72	-.30
Variance	.762	.997	.438	.596	.893	1.384	.743	1.456	.843	1.116
Skewness	-.85	-.64	-.62	-.02	-.12	-.48	-.34	-.43	-.20	-.18
UPPER										
Sum	19.452	20.938	14.670	15.082	13.886	18.887	16.399	19.823	19.382	18.576
Mean	1.945	2.094	1.467	1.508	1.389	1.889	1.640	1.982	1.938	1.858
StdDev	1.126	1.078	.712	.776	.678	1.087	.901	.985	1.031	.979
Kurtosis	-.38	.01	.67	.77	.58	-.33	-.31	.31	-.21	.47
Variance	1.267	1.161	.507	.603	.459	1.182	.812	.971	1.062	.958
Skewness	-.59	-.48	-.96	-.49	-.81	-.59	-.33	-.87	-.60	-.56

Table A.1: Sums of distances generated in two-dimensional Weighted MDS on the ordinal level, untieing ties, generated from upper and from lower triangle of input matrix

Test Statistics[a]

		B1	B2	B3	B4	B5	T1	T2	T3	T4	T5
Most Extreme Differences	Absolute	.133	.300	.300	.200	.300	.300	.167	.167	.333	.367
	Positive	.100	.300	.300	.167	.300	.067	.167	.167	.333	.367
	Negative	-.133	-.100	.000	-.200	-.067	-.300	-.133	-.033	-.133	-.033
Kolmogorov-Smirnov Z		.365	.822	.822	.548	.822	.822	.456	.456	.913	1.004
Asymp. Sig. (2-tailed)		.999	.509	.509	.925	.509	.509	.985	.985	.375	.266

a. Grouping Variable: NL

Table A.2: Non-significant results of Kolmogorov-Smirnov Z test between MOS of NSs and NNSs

Test Statistics[a]

		B1	B2	B3	B4	B5	T1	T2	T3	T4	T5
Most Extreme Differences	Absolute	.102	.024	.072	.047	.017	.052	.105	.183	.030	.042
	Positive	.000	.024	.072	.047	.017	.000	.105	.000	.000	.000
	Negative	-.102	.000	.000	.000	.000	-.052	.000	-.183	-.030	-.042
Kolmogorov-Smirnov Z		.880	.210	.620	.402	.144	.448	.903	1.581	.262	.362
Asymp. Sig. (2-tailed)		.421	1.000	.837	.997	1.000	.988	.388	.013	1.000	.999

a. Grouping Variable: NL

Table A.3: Kolmogorov-Smironv Z test comparing similarity judgements of NSs and NNSs

Test Statistics[a]

		stress
Most Extreme Differences	Absolute	.600
	Positive	.600
	Negative	.000
Kolmogorov-Smirnov Z		1.643
Asymp. Sig. (2-tailed)		.009

a. Grouping Variable: NL

Table A.4: Kolmogorov-Smironv Z test comparing stress values of NSs and NNSs

Descriptive Statistics[a]

	N	Range	Minimum	Maximum	Sum	Mean		Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
T1	30	6.6666E0	3.3333E0	1.00000E1	2.7999E2	9.3333E0	...	1.59019987E0	2.529	-2.981	.427	8.632	.833
B1	30	5.0000E0	5.0000E0	1.00000E1	2.7066E2	9.0222E0	...	1.46461019E0	2.145	-1.686	.427	1.916	.833
B4	30	7.0000E0	3.0000E0	1.00000E1	2.1899E2	7.2999E0	...	1.96784496E0	3.872	-.345	.427	-.804	.833
T5	30	7.6666E0	2.0000E0	9.66666E0	1.9866E2	6.6222E0	...	2.16508931E0	4.688	-.559	.427	-.841	.833
T4	30	7.6666E0	2.3333E0	1.00000E1	1.8733E2	6.2444E0	...	1.91772001E0	3.678	.149	.427	-.505	.833
T2	30	7.3333E0	1.6666E0	9.00000E0	1.6333E2	5.4444E0	...	1.87815294E0	3.527	.291	.427	-.460	.833
B3	30	8.0000E0	...	9.00000E0	1.2800E2	4.2666E0	...	1.74724742E0	3.053	.438	.427	.612	.833
B5	30	7.3333E0	...	8.33333E0	1.0933E2	3.6444E0	...	1.75279379E0	3.072	1.194	.427	1.719	.833
T3	30	8.0000E0	...	9.00000E0	9.4666E1	3.1555E0	...	1.76759536E0	3.124	1.132	.427	2.672	.833
B2	30	6.6666E0	...	7.33333E0	7.6333E1	2.5444E0	...	1.61526087E0	2.609	1.345	.427	2.239	.833
Valid N (listwise)	30												

Table A.5: Ranks of stimuli as generated from MOS task

Descriptive Statistics[a]

	N	Range	Minimum	Maximum	Sum	Mean		Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
S1	30	5.83	4.17	10.00	275.33	9.1778	.25496	1.39645	1.950	-2.646	.427	7.212	.833
S4	30	5.67	3.50	9.17	203.17	6.7722	.29852	1.63504	2.673	-.091	.427	-1.072	.833
S5	30	6.67	2.00	8.67	154.00	5.1333	.24622	1.34862	1.819	.221	.427	.800	.833
S2	30	6.83	1.33	8.17	119.83	3.9944	.27140	1.48652	2.210	.847	.427	1.368	.833
S3	30	6.50	1.50	8.00	111.33	3.7111	.25395	1.39095	1.935	.787	.427	1.723	.833
Valid N (listwise)	30												

Table A.6: Ranks of systems as generated from MOS task

References

Alvarez, Y. & Huckvale, M. (2002), The reliability of the ITU-T P. 85 standard for the evaluation of text-to-speech systems, *in* 'Seventh International Conference on Spoken Language Processing', ISCA.

Axmear, E., Reichle, J., Akamsaputra, M., Kohnert, K., Drager, K. & Sellnow, K. (2005), 'Synthesized speech intelligibility in sentences: a compariosn of monlingual english-speaking and bilingual children', *Language, Speech, and Hearing Services in School* **36**, 244–250.

Bennett, C. L. (2005), Large scale evaluation of corpus-based synthesizers: results and lessons from the blizzard challenge 2005, *in* 'Proc. Interspeech 2005', Lisbon, Portugal.

Benoit, C. & Grice, M. (1996), 'The sus test: a method for the assessment of text-to-speech intelligibility using semantically unpredictable sentences', *Speech Communication* **18**, 381–392.

Black, A. & Tokuda, K. (2005), The Blizzard Challenge-2005: Evaluating corpus-based speech synthesis on common datasets, *in* 'Ninth European Conference on Speech Communication and Technology', ISCA.

Borg, I. & Groenen, P. (2005), *Modern multidimensional scaling: Theory and applications*, Springer Verlag.

Bulut, M., Narayanan, S. & Syrdal, A. (2002), Expressive speech synthesis using a concatenative synthesizer, *in* 'Seventh International Conference on Spoken Language Processing', ISCA.

Campbell, N. (2005), Speech synthesis evaluation. PPT presentation, ELRA-HLT evaluation workshop, Malta.
URL: *http://www.elra.info/hltevaluationworkshop/img/pdf/Nick%20Campbell.ATR.Speech%20Synthesis%20Evaluation.pdf*

Clark, R. & Dusterhoff, K. (1999), Objective methods for evaluating synthetic intonation, *in* 'Sixth European Conference on Speech Communication and Technology', ISCA.

Coombs, C. H. (1964), *A theory of data*, Wiley.

Coxon, A., Jackson, J., Davies, P., Smith, H., Sachs, L. & Schmee, J. (1982), *User's guide to multidimensional scaling*, Heineman Education books.

Coxon, A. M. (2003), Multidimensional scaling, *in* M. Lewis-Beck, A. Bryman & T. Liao, eds, 'The Sage encyclopedia of social science research methods', Sage Publications, Inc.

Coxon, T. (2009), Multidimenstional scaling: 3-way analysis. PPT presentation.
URL: *http://www.tonycoxon.com/KUB/Module%207/INDSCAL-modified05.ppt*

de Cheveigne, A. & Kawahara, H. (2002), 'YIN, a fundamental frequency estimator for speech and music', *The Journal of the Acoustical Society of America* **111**, 1917.

Field, A. (2005), *Discovering statistics using SPSS*, SAGE Publications ltd, London UK.

Fraser, M. & King, S. (2007), The blizzard challenge 2007, *in* 'Proc. Blizzard Workshop (in Proc. SSW6)'.

Garofolo, J. et al. (1988), 'Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database', *National Institute of Standards and Technology (NIST), Gaithersburgh, MD* .

Gibbon, D., Mertins, I. & Moore, R. (2000), *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*, Kluwer Academic Publishers.

Hall, J. (2001), 'Application of multidimensional scaling to subjective evaluation of coded speech', *The Journal of the Acoustical Society of America* **110**, 2167.

Hirst, D., Rilliard, A. & Auberge, V. (1998), Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis, *in* 'The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis', ISCA.

Holmes, J. (2001), *Speech synthesis and recognition*, CRC.

House, A. S., Williams, C. E., Hecker, M. H. L. & Kryter, K. D. (1963), Psychoscoustic speech tests: A modified rhyme test, Technical report, U.S. Air Force System Command, Hanscom Field, Electronic Systems Division.

Jurafsky, D. & Martin, J. (2008), *Speech and language processing*, Prentice Hall.

Karaiskos, V., King, S., Clark, R. & Mayo, C. (2008), 'The Blizzard Challenge 2008'.
URL: *http://festvox.org/blizzard/bc2008/summary_Blizzard2008.pdf*

Malfrère, F., Dutoit, T. & Mertens, P. (1998), Fully automatic prosody generator for text-to-speech, *in* 'Fifth International Conference on Spoken Language Processing', ISCA.

Mayo, C., Clark, R. & King, S. (2005), Multidimensional scaling of listener responses to synthetic speech, *in* 'Ninth European Conference on Speech Communication and Technology', ISCA.

Meulman, J., Heiser, W. & SPSS, I. (2001), 'Categories 11.0', *Chicago: SPSS Inc* . **URL:** *http://www.courses.rochester.edu/SPSSDocs/SPSS%20Categorie%2011.0*

Montero, J., Gutierrez-Arriola, J., Palazuelos, S., Enriquez, E., Aguilera, S. & Pardo, J. (1998), Emotional speech synthesis: From speech database to TTS, *in* 'Fifth International Conference on Spoken Language Processing', ISCA.

P.85, I.-T. R. (1994), *A method for subjective performance assesment of the quality of speech output devices*, International Telecommunications Union publication. **URL:** *http://penta3.ufrgs.br/normasITU/HTML/P.HTMManual*

Podsiadlo, M. (2007), Large scale speech synthesis evaluation, Master's thesis, University of Edinburgh.

Pols, L. (1998), peech synthesis evaluation, *in* R. Cole, ed., 'Survey of the state of the art in human language processing', Giardini Editori e Stampatori, Pisa, pp. 429–430.

Raux, A. & Black, A. (2003), 'A unit selection approach to f0 modeling and its application to emphasis', *ASRU, St Thomas, US Virgin Islands* .

Sityaev, D., Knill, K. & Burrows, T. (2006), Comparison of the ITU-T P. 85 Standard to Other Methods for the Evaluation of Text-to-Speech Systems, *in* 'Ninth International Conference on Spoken Language Processing', ISCA.

Sluijter, A., Bosgoed, E., Kerkhoff, J., Meier, E., Rietveld, T., Sanderman, A., Swerts, M. & Terken, J. (1998), Evaluation of speech synthesis systems for Dutch in telecommunication applications, *in* 'The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis', ISCA.

Taylor, P. (2009), *Text-to-Speech Synthesis*, Cambridge University Press.

Turk, O., Schroder, M., Bozkurt, B. & Arslan, L. (2005), Voice quality interpolation for emotional text-to-speech synthesis, *in* 'Ninth European Conference on Speech Communication and Technology', ISCA.

Vainio, M., Jarvikivi, J., Werner, S., Volk, N. & Valikangas, J. (2002), Effect of prosodic naturalness on segmental acceptability in synthetic speech, *in* 'Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on', pp. 143–146.

Voiers, W., Sharpley, A. & Hehmsoth, C. (1975), 'Research on Diagnostic Evaluation of Speech Intelligibility.'.

Young, S., Odell, J., Ollason, D., Valtchev, V. & Woodland, P. (1995), 'The HTK book', *Cambridge University* **1996**.