Large Scale Speech Synthesis Evaluation

Monika Podsiadlo

supervisor: Dr Robert Clark

Master of Science in Speech and Language Processing

The University of Edinburgh

August 2007

Abstract

In speech synthesis evaluation, it is critical that we know what exactly affects the results of the evaluation rather than employing as vague notions as, say, "good quality speech". As so far we have only been able to rely on people's subjective judgements, a deep understanding of the mechanisms behind those judgements might lead to, first, an improvement in the overall quality of synthetic speech that could now address precisely the points that users find relevant, secondly, to a better subjective evaluation method design, and third, to designing an objective method of evaluation, that would offer stable and reliable comparison across and within systems. We base our work on the data from a large scale speech synthesis evaluation challenge, the Blizzard Challenge 2007, and we use a Multidimensional Scaling technique to map the acoustic features that users pay attention to while evaluating synthetic speech onto the dimensions along which the systems differ. We then present the results of a perceptual experiment conducted to test the hypothesis. The final parts of the thesis offer a discussion of the results and suggest some new directions for speech synthesis that follow from our findings.

Acknowledgements

First and foremost, the credit goes to my supervisor, Dr Rob Clark, for his advice and help, even with the silliest little things. And most importantly, thanks for being so kind and understanding. I really needed to work with someone who doesn't get upset over my crazy schedule. Otherwise, I would never have finished. Thanks.

My coursemates, Greta, Luca, John, Antonis, and Oliver- thanks for the coffees, the long hours in the lab, and generally for keeping me entertained and sane for the past twelve months.

I have also received lots of support and love from my family and friends back in Poland. Much appreciated, even if it was only over the phone. At the same time, I apologise for ignoring birthdays, e-mails, even Christmas, and for constantly putting off my visit. It's been a busy year, but I hope you agree it was worth it.

Then, I owe a lot to some people at IFA UAM- Dr Katarzyna Miechowicz-Mathiasen, Dr Jaroslaw Weckwerth, and Dr Jacob Thaisen. Jacob- thanks for introducing me to computational linguistics, for all the readings you suggested, and all the time you devoted to me during my final year at IFA. You got me here.

And finally, to Amit. For the "if it takes 8 hours, it takes 8 hours- you'll be in bed by 2", for answering my questions after 2 when I still wasn't, for putting up with me crying over arrays, and for making me laugh 40 hours and 8,000 words before the deadline. It was good to know you were always there.

Contents

List of figures

Chapter 1. Issues in speech synthesis evaluation

1.1 Introduction

Speech synthesis evaluation was until recently a somewhat neglected area of research, despite the fact that everyone generally agrees that there is a need for valid means of both comparison between different systems and within systems. After all, from the point of view of scientific research, without some sort of evaluation metric any real progress is virtually impossible. Thus, means of evaluating the performance of speech synthesis systems are needed to pinpoint their weaknesses and suggest changes that would possibly improve the performance and lead to further technological advances. The example set by the field of speech recognition (Black and Tokuda 2005) clearly shows that an evaluation measures can push the whole field forward. Even if the evaluation method seems rather simplistic and not very sophisticated, as is the case with a simple word error metric in speech recognition, it still provides some means of comparison between systems and a valuable insight into what research methods are worth pursuing further, and which the researchers would be better off abandoning.

1.1.2 Outline

The thesis begins with an overview of current issues in speech synthesis evaluation. Chapter 2 focuses on Multidimensional Scaling as a method of modelling human perceptual behaviour in speech synthesis evaluation tasks. Technical details of Multidimensional Scaling, as well as an overview of previous research using the technique, are provided. We then move on to an MDS analysis of a large amount of evaluation data obtained from the Blizzard Challenge 2007. The analysis results in

formulating a hypothesis that points to the acoustic features underlying people's judgements of naturalness of synthetic speech. Chapter 3 focuses on the perceptual experiment that we conducted to test the hypothesis, and examines its results. Chapter 4 offers a discussion of our findings in the light of current research in speech synthesis, and concludes.

1.2 Issues in speech synthesis evaluation

In the field of speech synthesis the matters seem less straightforward than in, say, speech recognition, where we know instantly what features we need to evaluate. In speech synthesis designing an evaluation method can be at times as complicated as designing a speech synthesis system itself, and many issues need to be addressed before we can actually move on to the evaluation itself. First of all, we are faced with a choice as to what we really want to evaluate. This decision will then guide the choice of an evaluation method. Thus, we might want to ask what qualities are the most important for the user when listening to synthetic speech, and choose to evaluate them as the most representative of the overall quality of the speech. We might definitely want to evaluate intelligibility, as the single most important factor contributing to the overall performance of the system. Intelligibility is a fairly straightforward feature, that can be measured with a reasonable accuracy by a Word Error Rate (WER). Word Error Rate is most commonly obtained from certain types of tasks that are typical for evaluating the intelligibility of synthetic speech - Semantically Unpredictable Sentences (SUS) or Modified Rhyme Test (MRT).

However, another common ambition of researchers is designing a synthesis system that would sound natural. We might then want to evaluate also the perceived naturalness of the synthesised speech. Unfortunately, characteristics such as naturalness are difficult to measure. To evaluate such a feature, we first need to determine what is understood by the notion of, say, naturalness, which is itself not an easy task at all, and also one requiring a careful analysis of human perceptual

mechanisms. Equally important, systems that achieve a very good Word Error Rate , but are not perceived as very natural-sounding, will definitely pose a problem. If we were to compare across systems, another metric would be needed to determine whether perhaps systems with poorer WER, but natural sounding are preferred over clear and intelligible, but robotic voices.

Another issue worth mentioning is what kind of evaluation we are interested in. For commercial systems, being subject to independent evaluation might not be beneficial at all. Poor results could naturally have a negative impact on the marketability of the product. Thus, certain parties might not be interested in developing a uniform evaluation metric at all, or they might want to design their own evaluation method, that would present their system more favourably. Another question is whether we want the system to be evaluated on its own, or in relation to other systems. It the latter, some standardizations are needed to ensure that such evaluation is justifiable. Given that today many voices tie a given database to a given domain, comparison between systems that are designed to serve different purposes might not be of much value either.

Finally, when evaluating synthetic speech we need to consider whether we want to rely on subjective or objective measures. Subjective measures rely on people's subjective judgements, while objective measures automatically evaluate the speech according to some predefined characteristics. Subjective evaluation is costly, as usually potential subjects expect to be rewarded for their participation. Also, even with such a financial incentive, it might be difficult to find enough subjects to validate the analysis (Bennett 2005). Then, people's judgements might not be very reliable, as many difficult to control factors come into play. Subjects might not be dedicated to contributing to scientific research, and only be motivated by the money paid in reward, which can potentially lead to careless judgements, and as a result meaningless to our analysis. Also, subjects might be tired at the time of the experiment, or can stop paying enough attention at some point due to fatigue. They might not understand the tasks completely, but be too embarrassed or uninterested to admit and seek further

explanation. All those elements affect the results which we then base our analysis on. Thus, the quality of the whole analysis is largely dependent on factors we are not able to control.

Naturally, an alternative would be to rely on objective measures. However, attempts at designing such a method have not been very successful so far, and the correlations between the results of such evaluation with standard perceptual tests are still inadequate. Clark and Dusterhoff (1999) report such a lack of correlation when testing objective methods of evaluating synthetic intonation. Also, objective evaluation methods have to be evaluated as well, and this is most commonly done by correlating the results with the results of perceptual experiments. This means that we are again relying on subjective methods of evaluation. This leads to a situation when we have to ask whether it is the evaluation method that does not model human behaviour accurately, or the quality of our data from perceptual experiments is inadequate. All in all, we are currently not able to construct an objective method of evaluation with satisfying results, and need to rely on subjective methods. If any progress towards designing a reliable objective method is to be made, we need a better understanding of subjective evaluation methods as well as human perceptual mechanisms.

1.3 Subjective methods of speech synthesis evaluation

In subjective methods of evaluation one of the most commonly used tests are Mean Opinion Score tests. In MOS tests subjects evaluate a given feature of synthetic speech on a predefined scale, usually 1-5. The study by Toda et al (2002) presents one of the applications of MOS in speech synthesis. Here, Toda et al work with the idea that efficient cost functions need to reflect perceptual characteristics. Thus, they ttest if the cost functions employed in the engine model accurately human perceptions of speech. That is, if the increasing cost is accompanied by a decreasing level of perceived naturalness. To do so, MOS is used to elicit judgements from subjects upon

presentation of a stimulus representing different cost calculations.

Another popular method in subjective evaluation are Semantically Unpredictable Sentences tests. SUS tests (Benoit et al 1996) stem from the observation that meaningful sentences provide the listener with contextual cues that contribute to the perceived intelligibility of synthesised speech, and thus distort the results and are not reliable as an evaluation method. Another observation here is that most existing speech synthesis systems can already handle meaningful sentences quite well. Thus, it is only on the basis of those meaningless ones that we can evaluate and compare across different systems in any meaningful way.

The SUS method consists in generating at random a set of sentences that are syntactically well-formed, but semantically anomalous, synthesising them and performing a listening test. Thus, we have a set of structurally basic sentences in which each word's subcategorization frame is represented, although the specific choice of words is very random. The SUS test is equally useful when both evaluating just one synthesiser or evaluating and comparing more systems.

Compared to other evaluation methods (Benoit et al 1996), SUS sentences are the most challenging tests for a speech synthesiser, with intelligibility scores ranging from 10% to 20%. Higher scores were attained using methods that provided contextual cues. This shows that indeed, removing contextual information from the input stratifies synthesis systems according to their performance more accurately.

Another interesting evaluation method (Edgington 1997) could be testing the performance of a speech synthesis system by assigning it a task that lies just beyond its expected capabilities. Such tests can thus be an excellent challenge identifying the really good systems. Edgington (1997) present three tasks for evaluating concatenative synthesis systems that are slightly beyond the scope of the system. The tasks include emotional synthesis, synthesising glottalization, and experiments with speech database.

The emotional synthesis task involves collecting prosodic data and trying to map it onto the synthesised speech. This is done by collecting a sample of speech from a

voice talent who says each of the sentences in a particular emotional style. Then the prosodic features of these utterances are identified, and mapped onto a speech synthesised using one of the concatenative systems. Thus, we have two sets of identical sentences expressing different emotions, one set with synthesised utterances, and one with natural speech. A perceptual experiment is then conducted, where subjects are asked to match each sentence in both sets with a perceived emotional style. The results show that the average recognition rate for natural speech is 79.3%, while for synthesised speech only 42.2.% This in turn suggests that including some more sophisticated prosodic features in synthesis that would be clear and recognizable is still a long way off.

The next suggested task, the glottalization task, consists in simulating the effect by manipulating F0, duration and amplitude. In previous approaches F0 would simply be lowered to 35/50 Hz for a duration of 100ms. The resulting sound would be highly unnatural, so to rectify the situation the duration of the preceding sound would be extended by 100ms. The result was still unsatisfactory. The new approach to time-domain synthesis for glottalization, on the basis of the analysis of the glottalization database, suggests that F0 be lowered to 70Hz in only 20ms preceding the glottalization, followed by a 70ms silence. In a listening test, where subjects were asked to evaluate which synthesised utterance resembles the natural one best, the one synthesised using the modified approach scored best. Also, another experiment was conducted to test if the modified approach is of any help when dealing with cases where glottalization causes ambiguity. Subjects were presented with three sets of ambiguous expressions: one of glottalized, one of unglottalized, and one of synthesised utterances. For the synthetic speech, intelligibility rate ranges from 86% to 96%, which proves the modified model very effective .

Considering the nature of the speech database, the choice is between diphones, demi-syllables, and phonetically balanced continuous speech. Experiments show that intelligibility and naturalness is not directly linked to the type of the database, but also depends on the type of output. If we wantto synthesise single words, demi-syllable

database is most effective. In case of whole sentences, however, continuous speech database performs better than the other two, that caused the sentences to sound overarticulated (Edgington 1997).

Another method of subjective evaluation in speech synthesis is the Modified Rhyme Test (House et al 1963). Modified Rhyme Test is based on the use of a carrier phrase, e.g. “Now we will say __ again” (after Bennet 1997), which subjects then have to complete with the word they hear. For this task, a large list of monosyllabic words is constructed, so that each word from the list can be confused with at least a few others on that list. These words are then used in a listening experiment, and the participants have to complete the carrier phrase with the word from the list they hear. The MRT serves as a method of testing the intelligibility of synthesis systems similar to Semantically Unpredictable Sentences.

The methods described above were all either methods manipulating the speech, so that it displays certain characteristics, or methods with which those characteristics can be later evaluated with. The Blizzard Challenge employs the latter to test how different speech synthesis systems perform on a uniform dataset. The results from those evaluation tasks will serve as a basis of our analysis.

Chapter 2. The analysis

We want to conduct a comprehensive study of people's responses to discover what acoustic features underlie their judgements of synthetic speech. This will help us to understand what aspects contribute most to the perceived naturalness of synthetic speech. This understanding, on the other hand, can lead to a better subjective evaluation tasks design, and finally, to creating an objective method of evaluation of synthetic speech that would model human perceptions accurately, while being free of their limitations. To do so we need a large amount of evaluation data. This is obtained from The Blizzard Challenge 2007 . Both the analysis and the evaluation part of the research are based entirely on the data from the challenge.

2.1 The Blizzard Challenge 2007

The Blizzard Challenge (Bennett 2005) is a large scale speech synthesis evaluation project, that was designed as a way of comparing different synthesis techniques and a forum for exchanging ideas and research methods. The main idea behind the challenge is that in a field with so much variance, where the quality of synthetic speech is so largely dependent on databases and the domain in which the system is used, we can only aim at comparison across systems if the same database is used for the same tasks. Thus, Blizzard challenges research groups and companies to build a voice from a provided database, and then synthesise a prescribed set of sentences that are then evaluated (Black et al 2006). The project started in 2005, and has gradually attracted more and more participants. Despite the initial concerns that commercial systems might not want to be subject to independent evaluation, while in 2005 only six systems participated, the 2007 edition saw as many as sixteen systems taking part, both commercial and academic. Participation in the challenge is anonymous, and systems are assigned an identification letter. After the initial results of the challenge were released, the participants were asked to fill in a questionnaire, describing the technology used in the synthesiser and providing general feedback on the whole challenge. The feedback was clearly positive, with most participants declaring to participate in the forthcoming edition.

The database for the 2007 challenge was released by ART-SCL, and comprised an 8-hour recording of male American English voice. The participants were asked to submit a voice using the full released database (voice A), using the CMU Arctic database (voice B), and using a chosen subset of the full database. Sixteen systems submitted voice A and B. Eleven systems submitted voice C. The participants had a month for training and a week for the actual synthesis. Then, the sentences synthesised by each system were put for a web-based evaluation. In addition, "system" I was included in the evaluation- the set of sentences recorded by the original speaker. This served as a reference point and enabled a direct comparison not

only between synthetic speech, but also of synthetic speech with natural voice.

The speech was evaluated by different groups of users to ensure as wide coverage as possible. Thus, the users were both native and non-native speakers of English, with different levels of experience in listening to synthetic speech, from different backgrounds and age groups, some of them had an extra incentive of being paid in reward for their participation, others had no such motivation. Obviously, one might question the validity of any analysis based on data collected over the Internet, where there is virtually no means of controlling participants responses. However, the Blizzard organisers feel that the amount of data obtained from web-based evaluation can be a lot greater than if any other method were to be used, and thus offers a more comprehensive analysis. This, it is believed, offsets any potential noise in the data. All in all, four groups of users participated in the evaluation. First, paid UK students (native speakers of British English), paid US students (native speakers of American English). Next, a group of speech experts, and finally, volunteers, i.e. anyone who was willing to participate and registered on the website. The evaluation was conducted over the Internet and scheduled for one month. Upon completing all tasks, users were asked to optionally fill in a questionnaire, providing some demographic information as well as feedback on each section of the evaluation.

This year's evaluation consisted of five tasks. In task 1 users judged the similarity of synthetic speech to natural speech. They first listened to four samples of the original voice, and then to one sample of synthetic speech. They were then asked to rate the similarity of the synthetic speech to the natural voice on a scale from 1 (sounds like a totally different person) to 5 (sounds exactly like the same person). The results of this section form the "similarity" scores we are going to refer to in subsequent sections. In task 2, users heard pairs of samples from different systems (or, in the case of system ordering, from the same system but a different dataset). They were asked to judge how similar in terms of naturalness the two samples sound, ignoring the meaning of the sentences and just focusing on their naturalness. The results of this section will be of the most interest to us, as they will be used as the

basis for the whole MDS analysis. Tasks 3 and 4 were Mean Opinion Score (MOS) tasks, where users heard synthetic sentences from the conversational and the news domain respectively. They were then asked to rate how natural the sentences sound on a scale from 1 to 5. The MOS scores we use in the analysis come from both sections. The last task was a Semantically Unpredictable Sentences test, which we described in detail in previous sections. Users were asked to transcribe what they heard without the aid of contextual information. The task tested the intelligibility of speech synthesised by the participating systems, and its results form the Word Error Rate (WER) scores that we use later in the analysis as well.

The results of section 2 present "similar" and "different" scores for pairs of systems, where a pair B_A does not necessarily score the same as a pair A_B. We convert the scores from section 2 for voice A automatically into a dissimilarity matrix, dividing the number of the "different" scores by the total. We then input the data matrix into SPSS 14.0 and run MDS PROXSCAL Version 1.0.

In addition to the similarity judgements, the MDS analysis was also based on questionnaires submitted by the participants, as well as samples of synthetic speech from the news domain. In most cases the questionnaires offered a fairly detailed overview of the systems and the technology behind them. We also feel that the speech samples were comprehensive and representative of the overall quality of the system, even though they did not cover, for example, conversational domain. Still, sentences of varied length and vocabulary, including proper names, were included, giving us a fairly good idea of what subjects heard in the original evaluation tasks.

2.2 Multidimensional Scaling.

Multidimensional Scaling, which we base our analysis on, is a technique of representing similarity or dissimilarity judgements among pairs of objects as distances

on a low-dimensional, multidimensional plane (Borg and Groenen 2005: 3). The data is mapped onto a graph with points representing objects, and distances between the points representing similarity/dissimilarity judgements between the respective objects. As MDS aims to be visually as clear and straightforward as possible, the scaling is most commonly performed in Euclidean geometry. Such a graphical representation is a lot more straightforward way of exploring the data than simply trying to make sense of numbers, and allows the researcher to observe regularities that might otherwise go unnoticed. Thus, Multidimensional Scaling provides a powerful visual tool for data analysis.

In more technical terms (Borg and Groenen 2005: 37-42), Multidimensional Scaling transforms "proximities p_{ij} into distances on an m-dimensional MDS configuration X by a representation function $f(p_{ij})$ specifying relationships between proximities and distances $d_{ij}(X)$." Rather than satisfying f in full, however, a model that approximates f as closely as possible is sought, with the level of this approximation defined as the loss function, most common of which is stress. Stress is a normed sum-of-squares of representation errors $e_{ij} = f(p_{ij}) - d_{ij}(X)$, over all pairs (i, j). The actual MDS representation is created by defining a set of m axes, perpendicular to each other and intersecting in one point, which define an m-dimensional space. Then, each point in that space can be defined by an m-tuple $(x_i1, x_i2, \ldots, xim)$, where x_ia is a projection of i's onto the dimension a. Euclidean distance between pairs of such points can be then computed using the Pythogorea
theorem (for a 2-dimensional representation):

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

The above formula can be further rewritten for any given number of dimensions.

Multidimensional Scaling can be used for different purposes. Borg and Groenen (2005) define four most common uses of MDS. First, MDS can be used as an exploratory technique on data that does not display any obvious correlations. This

applications is aimed at disclosing patterns and structures in seemingly random data. Secondly, MDS can be used for testing structural hypotheses when more is known about the data. Then, Multidimensional Scaling can be used in psychology for exploring psychological dimensions underlying similarity or dissimilarity judgements. And finally, MDS can serve as a model of similarity judgements, when it is assumed that the computed distances between the objects reflect "psychological" distances between those objects as perceived by the subject. This last use of MDS will be of most interest to us, as our analysis aims at providing a model of human behaviour when listening to synthetic speech.

2.2.1 Technical considerations

Multidimensional Scaling can be an excellent tool provided that it is used correctly. When modelling similarity judgements, it is important that we know what kind of data we are working with, how to use it, and how to interpret the results.

As mentioned before, MDS analysis most commonly uses Euclidean distance. Irrespective of the type of representation, in metric geometries (i. e. geometries that allow one to measure the distances between its points) the notion of distance has a few properties that are important for MDS. There are three definitional characteristics that are always true of distances, and that need to be satisfied for any function assigning values to distances to hold. These three distance axioms are nonnegativity, symmetry, and triangle inequity (Borg and Groenen 2005: 33-34). Nonnegativity of the distance function consists in the distance between any two points *i* and *j* being greater than 0 or equal to 0 (when $i = j$):

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

Symmetry is the property of the distance so that for any given points i and j the distance between *i* and *j* is the same as the distance between *j* and *i*:

$$d_{ij} = d_{ji}$$

Finally, triangle inequality says that for any three points i, j and k the distance between i and j will never be greater than the sum of distances between i and some intermediate point k, and k and j:

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

Those properties are important for constructing MDS representations, as they ensure that the proximities can be transformed into distances in geometrical sense. If the proximities we base the analysis on do not satisfy those axioms, the data might not be valid for any MDS analysis.

MDS can be used on both ordinal and ratio data. For ratio data, the ratios correspond to the ratios of the distances between the objects on the graph. For ordinal data, we are only interested if the order of objects along the dimensions computed by MDS reflects the order of objects in the data. The data we are scaling is ratio (Stevens 1946). This gives us a range of admissable transformations that can make the analysis more easily readable (Borg and Groenen 2005: 23). For ratio MDS, we are allowed to perform a set of *rigid motions* or *isometric transformations*, that can make the representation more readable, yet leave the distances invariant, i. e. we can transform the configuration by rotating or reflecting it, as well as translating (changing the configuration relative to a chosen point) without affecting the distances. Such isometric transformation will be performed on our data to adjust the axes so that the natural speech serves as a reference point for all other objects.

The most potentially challenging aspect of an MDS analysis, however, is choosing the *m* value – specifying the number of dimensions. Most commonly, we try to arrive at as low-dimensional a representation as possible (Borg and Groenen 2005: 65), so as to smooth the representation of possible over- and under-generations due to error in the data. Also, lower-dimensional representations are easier to analyse initially, and if

there is such need, the number of dimensions can be increased once we gain a better understanding of the data.

Generally, deciding on the number of dimensions, we need to evaluate how well the configuration reflects the data. However, as the proximities we use are rarely perfect (Borg and Groenen 2005: 41-42), containing certain imprecisions, we want our representation to only approximate the data, rather than representing it exactly, to allow for smoothing of those imprecisions. Paradoxically, those imprecise approximations in practice reflect the proximities better. Evaluating this badness-of-fit, we most commonly look at stress. The stress value can be both informative and confusing at the same time. On the one hand, there are some rough rules for evaluating stress, like, say, any stress below the value of 0.20 is acceptable and signifies a good enough fit. On the other hand, such rough rules do not always hold, and what we think is a meaningful stress value can be just an artifact of, say, the missing values, the error in the data, or the number of objects to the number of dimensions ratio (in principle, the fewer the objects and the higher the number of dimensions, the closer to 0 stress we can get).

One way of looking at stress is obtaining stress values for most likely *m*-values, and using the one which marks the last stage of any real improvement (Borg and Groenen 2005: 47-49), meaning that the further increase does not improve the stress significantly. This can be done looking at the scree plot, a curve connecting points with (*m*-value, stress-value) vectors on a 2-dimensional diagram, and trying to locate the "elbow" -a point from which onwards the decrease in stress is less significant. The *m*-value that the elbow marks is the one that should be used.

Finally, we might want to choose the number of dimensions that make most sense to us, and that fit in well with our predictions as to the data.

2.2.2 Previous research on MDS in speech synthesis evaluation

Ours is not the first attempt to use Multidimensional Scaling in modelling human behaviour in speech synthesis evaluation. Mayo et al (2005) tested Multidimensional Scaling as a possible technique in determining what acoustic features users pay most attention to. In a pilot study Mayo et al (2005) first synthesised a set of 8 sentences of varying degrees of perceived naturalness using Festival Multisyn. Eight adult native speakers of English participated. The task was very similar to task 2 used in the Blizzard Challenge 2007. The participants heard pairs of sentences and were asked to make a simple binary decision on each pair- whether the two sentences were similar or different in terms of their naturalness. They were asked not to focus on any particular features of the speech, like intonation or joins, just on the perceived degree of naturalness. The participants also had a chance to hear a few pairs of samples of synthetic speech representing extreme cases of similarity or dissimilarity in terms of naturalness, so as to familiarize themselves with synthetic speech and understand the task better. A pre-task testing this understanding was also administered. The results of the experiment were then converted into a dissimilarity matrix and subject to an MDS analysis.

Mayo et al (2005) report that the 3-dimensional configuration offered the best fit, with residual stress at the 0.05004 level. Their analysis suggests that users arranged the utterances on a graded scale, with natural examples on one end and unnatural ones on the other. The visual representation also revealed that the sentences were grouped into three clusters: one with fairly natural sounding sentences, another with sentences that had rather serious errors in prosody (duration, or intonation, or both), and the last one with sentences displayed errors on segmental level: bad joins, resulting from poor unit selection. In addition, some sentences that combined to an extent characteristics of more than one cluster, were placed in between those clusters, supporting the analysis (Mayo et al 2005).

Mayo et al conclude that MDS is indeed a valuable tool for discovering the

dimensions that underlie people's judgements. The study suggested that people make judgements about naturalness along at least two dimensions: the appropriateness of prosody, and the appropriateness of units used in the synthesis. Also, it is stressed that making simple binary decisions about the perceived similarity of pairs of samples is a fairly straightforward task that participants had no trouble completing. However, the results of this simple task explain a great deal about as complex processes as auditory perceptions.

Our study to a large extent replicates the one by Mayo et al (2005). The data we analyse also comes from a task where subjects were asked to state whether pairs of sentences are similar or different in terms of their perceived naturalness. However, our study offers a more comprehensive analysis, as we are working with the data from a larger evaluation task. While we do not doubt that samples used by Mayo et al presented different degrees of naturalness, we note that they were all synthesised using one speech synthesis system. Thus, the analysis might not be valid for systems employing a totally different technology to the one used in Festival Multisyn[1]. Our study, on the other hand, extends the previous one in two ways. First, we analyse a wider range of data obtained from more subjects evaluating more and more varied samples. Secondly, we conduct another perceptual experiment, this time testing the hypothesis formulated on the basis of the MDS analysis. The second experiment serves to either confirm or disprove the validity of the proposed MDS analysis, in that we decide to ask the participants specific questions rating different speech synthesis systems according to the features identified as representing the dimensions from the MDS configuration. Thus, we offer a fuller discussion of both whether Multidimensional Scaling can model well human perceptual behaviour in speech synthesis evaluation, as well as weighting of different acoustic features by users evaluating synthetic speech.

[1] Festival Multisyn (Clark et al 2007) is a unit selection engine based on diphone-sized units only and direct feature-based target cost for determining the suitability of a given unit.

2.3 Running MDS over the Blizzard data

To perform an MDS analysis of the Blizzard data we use SPSS PROXSCAL We first input the dissimilarity matrix we created from the dissimilarity judgements into SPSS. As our data is ratio, we select the ratio analysis using the full matrix. We mentioned earlier that symmetry in the data is vital, yet we also mentioned that in the matrix a pair A_B does not correspond to a pair B_A. When the "full matrix" option is used in SPSS, if the matrix is asymmetrical, the values are weighted and the matrix symmetrized.

2.3.1 Choosing the number of dimensions

As we noted before, one of the biggest difficulties of Multidimensional Scaling is choosing the most appropriate number of dimensions. It is especially challenging in a situation when we not only want to confirm our predictions, but we need to construct a hypothesis from scratch, with no expectations as to the number of dimensions, or the features that the dimensions represent. Thus, we begin by testing how MDS works on 2, 3 and 4 dimensions. We evaluate the badness-of-fit measures, looking at stress, the scree plot and the transformation and the residual plot, which are another diagnostic techniques to determine how well the proximities reflect the original data. The stress values are 0.05720, 0.4584, and 0.03668 for 2, 3, and 4 dimensions respectively. The scree plot displays a clear "elbow", suggesting a 3-dimensional view as the most appropriate. However, we decide on a different line of reasoning, and choose the 2-dimensional view as the best fit. Looking at just the values, not their graphical representation, we notice that there is no real improvement in the fit with the higher dimensions. as we are only dealing with differences of approximately 0.01. At 0.05720 for two dimensions, the stress is already very low, indicating a very good fit. Although it does drop further to 0.04584 for 3 dimensions, and even further to

0.03668 when 4 dimensions are used, we consider this to be the usual decline that comes with increasing the number of dimensions, and therefore insignificant.

The most powerful argument, however, comes from the analysis of coordinates after adjusting the axes. Using SPSS, we correlate the new coordinates along both dimensions with WER, MOS and similarity to the original speaker scores. The results show that there are significant correlations between the coordinates and most scores. Only dimension 1 does not seem to be significantly correlated with WER[1]. However, the lack of correlation here is understandable and expected.

Also, we tried listening to the samples in the order suggested by the graph. Such an auditory analysis by an investigator might seem extremely subjective and thus unreliable. However, it is a standard procedure in deciphering MDS configurations, and therefore should not be dismissed.

To our ears, the positioning of the systems on the graph corresponds clearly to two acoustic dimensions, rather than three or four. Grouping the samples so that they differed by one dimension only, and listening to them along those continua revealed a pattern that fitted well into the 2-dimensional analysis. Also, our perceptions were supported by the information from the questionnaires. Certain regularities were observed that reflected the technology used in a given system. The expected quality of speech when a given technology is used matched well our predictions as to the dimensions and the features they represent.

Finally, the 2-dimensional view seemed visually rather obvious to analyse. The objects in the plane seemed to be placed along exactly two dimensions- at this stage not really corresponding to the axes. The objects formed a rather clear square configuration, with system I-the natural voice, occupying the bottom-left corner of that square, and other systems positioned along what seems like axes perpendicular to horizontal axis just above system I.

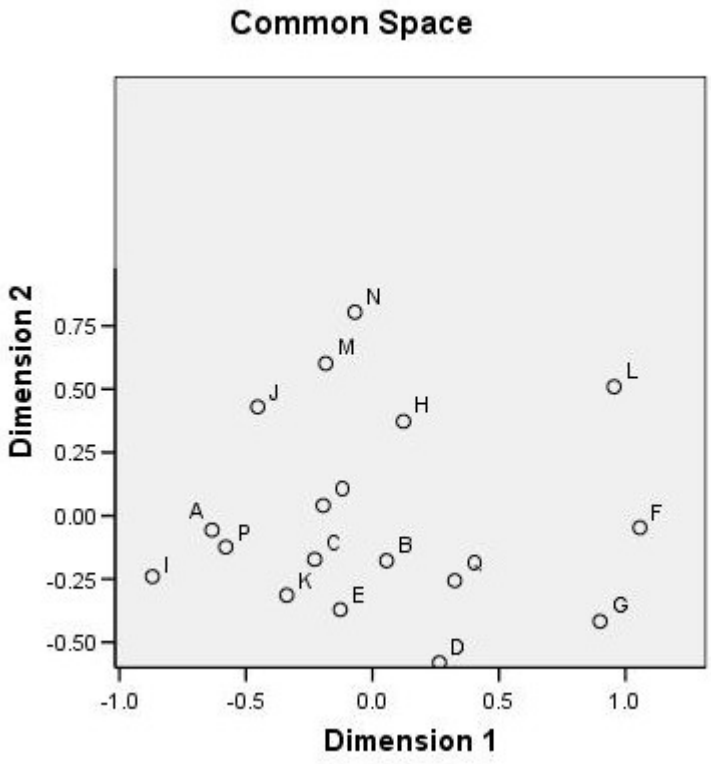
Having decided on the number of dimensions, we arrived at a graphical

---

[1] The reason for the lack of correlation here will become apparent once we map acoustic features onto the dimensions. We will see that the feature chosen should logically only have little or no effect whatsoever on the WER scores.

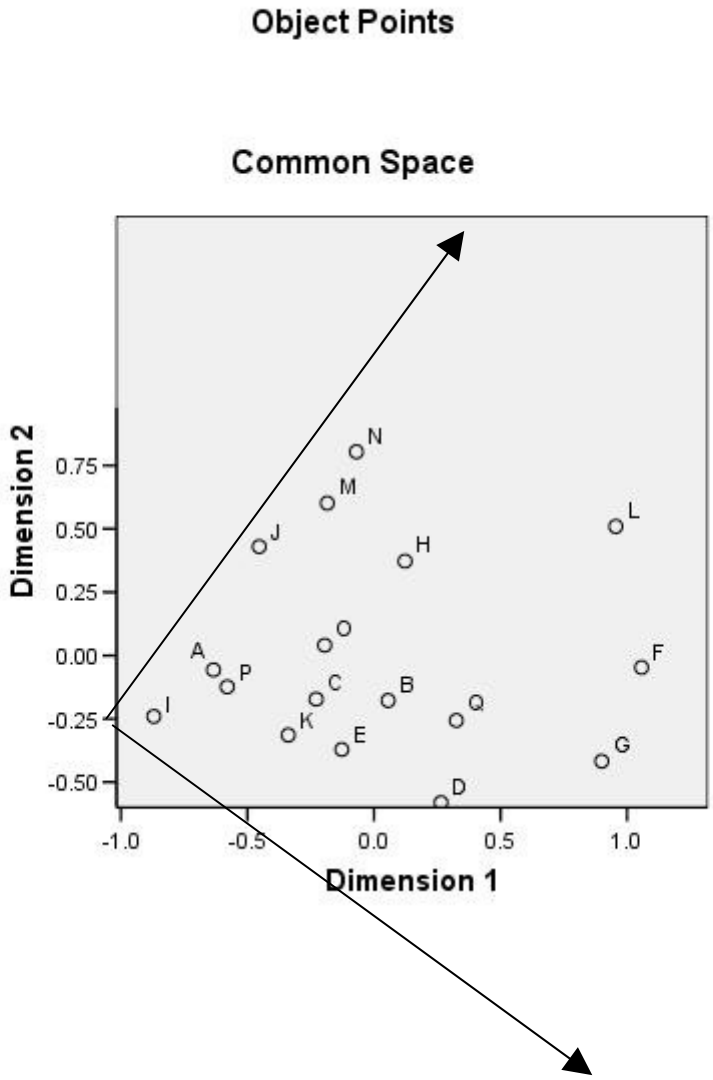
representation of systems similarities:

Figure 1. The original 2-dimensional view.



The positioning of the axes, however, is quite arbitrary here, and at this initial stage any potential regularities are difficult to observe. Since isomorphic transformations are allowed with ratio MDS, we rotate the graph to arrive at the clearest representation. To do this, we assume that I- the natural speech- serves as some sort of ideal that other systems aspire to. That this is true is also supported by the results of all tasks used in the Blizzard, where system I was consistently rated as the best, without the users knowing that they are dealing with natural speech. Therefore we decide to position the axes so that I is farthest down and left:

Figure 2. The rotated graph



With axes positioned in this way, we get a clear 2-dimensional view. This new graph serves as the basis for all further analysis.

2.3.2 Analysing the dimensions

To analyse the dimensions, we select groups along the axes, then compare the ordering of systems within those groups against the information from the questionnaires and the samples of synthetic speech. First, we select groups along the horizontal axis, as graphically they form the most obvious regularities. Thus, the leftmost group would include systems P, A, J, M and N. Those systems approximate best the natural voice in terms of Dimension 1. Next, systems K, C, O and H could form group 2. Then, systems E and B seem to be positioned along approximately the same line,. The same goes for systems D, Q, and L, forming our fourth group, and system F and G as the last group, right-most on the horizontal axis.

A similar grouping can be performed along the other dimension- the vertical axis. Starting from the bottom, we have the systems that approximate the natural voice best along Dimension 2: systems A, P, K, C, E and D. Then, slightly higher up, systems J, O, B, Q and G. Then, systems M, H, and F, with systems N and L forming the last group. Those groups are just indicative of the general tendency and should be treated as such, as there is no real reason why, say, system C should belong with E and K rather than O and B.

After listening to the samples of speech in the order suggested by the groups, so that the samples within the groups differ by approximately one dimension only, and analysing our perceptions in relation to the questionnaires, we formulate a hypothesis that the higher the system on the vertical axis, the more robotic the speech, and the farther to the left on the horizontal axis- the worse the joins. This explains why there is no correlation between the WER and Dimension 1. The voice does not have to sound human to be clear and intelligible, and more robotic voices can still achieve a low WER, as long as they stay clear. And the other way round, a human-like voice does not guarantee a good WER if there are other problems, like, say, bad joins. Thus, the lack of correlation here is understandable.

2.3.2.1 System profiles and the analysis

Looking at the graph more closely, along the dimension 1 in the first group we can see that systems achieved a very good quality of joins. System P is a concatenative system that produces very high quality smooth speech, attributed by the developers to their Unit Selection with Low Time Scale Modification. The samples that we listened to were rather impressive, with a very natural voice, few audible joins, and very good segment quality. System A also produced very smooth speech, with very few audible joins, and in addition, with very accurate and quite animated intonation. This could be the result of F0 and duration HMM-based unit selection, and probabilistic join and target cost calculations. However, in relation to system P, speech synthesised by system A sounded more buzzy and slightly less human-like, although with more natural intonation. System J, that was placed higher up on the axis, offers HMM-based parametric synthesis. According to both the graph and our auditory analysis, this is the system that has least, if any, audible joins. However, the developers point to "very muffled speech quality caused by the parametric synthesiser" as the weakest point of the system. This is completely confirmed by our analysis and the place of system J along the vertical axis. All we know about system M is that it is concatenative. The speech synthesised with system M is of overall good quality, very consistent and clear, with slightly audible joins, but in a way that does not affect intelligibility. which is confirmed by the WER score. However, the speech sounds robotic, and in this respect not very natural. System N is an HMM-based synthesiser using context-dependent quinphones, offers very smooth speech with no audible joins, however, the resulting speech sounds very robotic. The developers quote the stochastic model used in the system as the reason for low similarity to the original speaker score. This group is the one where the systems are configured in such a way that most closely resembles our ideal model, with objects differing by one dimension only. Thus, it is probably easiest to confirm our hypothesis about Dimension 2 corresponding to the perceived "roboticness" of speech looking at this very group.

Farther to the right along Dimension 1 we have a group with four concatenative systems. System K uses phones and half phones, and profits greatly from HMM-based speech recognition based automatic labelling. The quality of speech is similar to that of system A, however system K is less consistent in segment quality. Otherwise it would probably be more to the left on the horizontal axis, as at times the quality of speech is excellent, with no audible joins. At other times, however, the samples display problems with segment duration and awkward prosody, which results in a choppy voice at times.

System C is another concatenative systems that primarily uses diphones, and occasionally relying on halfphones. The system offers good quality speech with occasional problems on the segmental level, resulting in poor joins, that do not affect the intelligibility of the speech. Rather, similarly to system K, it sometimes sounds a bit awkward. We do not know much about system O, except that it is concatentive. The voice is clearly more robotic than the previous two, but still good quality is maintained. There are occasional problems with joins .The system is not consistent here, as there are times when the speech sounds smooth with few audible joins. System H, another diphone-based concatenative system, sounds more robotic and joins at times affect intelligibility. However, this is the one system we feel that perhaps we have rather unrepresentative samples of, as the WER, MOS and Similarity scores are better than we would have expected judging from the samples only. All in all, this group includes some decent systems that offer good quality speech, but are inconsistent at times, and occasionally display segmental problems. The roboticness, again, varies, and reflects well the position of the objects in the MDS plane.

The third group along Dimension 1 includes two concatenative systems. System B uses diphones, and the resulting speech is very consistent and of decent quality. However, the joins are audible most of the time. They usually do not affect the intelligibility, but they are there, making the speech sound less natural. System E sounds very similar to be., but is a bit more consistent and human-like.

The next group contains two concatenative systems, and one system combining concatenation with statistic parametric synthesis. The latter, system D, has rather audible joins and flat intonation. There are some problems on segmental level. However, other than these problems, the sounds is quite similar to the original speaker. The developers admit that the greatest problem with the system are the joins due to insufficient smoothing, and this is reflected by MDS. System Q uses context-dependent demiphones, and was originally designed for a language other than English. This all results in rather robotic, very choppy speech, which is still intelligible, though (WER = 0.32). System L, on the other hand, is a concatenative system based on demiphones, with ASR-based unit selection. The resulting voice is extremely robotic and very choppy, with a lower WER (0.47), but here it is rather the roboticity than the joins that affect the intelligibility. This extreme roboticness is in line with our analysis, as system L was the was placed the highest along the vertical axis.

Finally, the last group includes two system with very bad joins and poor intelligibility. System G is, again, a concatenative system with variable-sized unit selection. It has been under development for only three months, so the poor scores are understandable. This is the system that had the poorest WER (0.81), most probably as a result of the worst joins. This is also reflected by the right-most position in the MDS configuration. System F had slightly better joins, although the intelligibility with the WER of 0.61 was still very poor. Speech synthesised by system F was, unlike G, very muffled and robotic. System F is a concatenative system with diphone-sized units. It was mentioned in the questionnaire that the system's performance was greatly affected by the choice of a lexicon (Celex), and bad forced alignment acoustic model, and that an improvement was noticed when CMUDICT was used instead.

In short, the MDS configuration corresponds well to both our auditory perceptions and the information from the questionnaires. It also shows that although the two dimensions underlie people's judgements of perceived naturalness, it is the Dimension 1- the quality of segments, that affects the performance most- by having a seriously

detrimental effect on the intelligibility of synthesised speech.

Also, the analysis is based on the results of a task where users were asked to decide whether the two systems they heard samples of were similar or different in terms of their naturalness. Thus, the users based their judgements of naturalness on the two features we mapped onto the dimensions. The users, however, were not asked which voices they preferred, or which voices they would rather listen to. Therefore, the analysis does not answer the question of what makes a successful or likeable system, unless by those notions we mean natural. The analysis, however, shows that the two things most detrimental to the perceived naturalness of the voice are bad joins and roboticness. To confirm or disprove those findings, a perceptual experiment was conducted.

Chapter 3 Testing the hypothesis

Since we have identified two dimensions along which the systems were differentiated by the users, we now try to confirm if we have done so correctly, and if indeed the two suggested features were what the similarity judgements were based on. To test the hypothesis, we conduct a listening experiment. To avoid the controversy over whether non-native subjects are able to make stable and meaningful judgements, we decide to limit ourselves to native speakers of English. Participants are paid £5 each in return. In terms of the stimuli, we decide to use the samples submitted for the Blizzard Challenge. We select one sentence from newspaper text, as it appears to be the most representative from all sample sentences that were made available to us. The sentence reads: "But the wrench thrown by Golisano into the conventional wisdom has clearly increased the stakes in the West." The sentence is just about long enough, and seems to reflect reasonably well the overall quality of speech synthesised by a given system.

Since to test the hypothesis best we need stimuli that would clearly represent the feature in question, we group systems as they appear on the graph, so that the test

samples differ approximately along one dimension only. Thus, testing the first dimension, we arrive at five groups of samples that cover all sixteen systems: the first being systems A, J, N, M and P, the second: C, H, K and O, the third: B and E, the fourth: D, L, Q, and the fifth systems F and G. This gives a total of sixteen approximately 8-second samples in the first part of the experiment. Testing the second dimension, we used samples from fourteen systems divided into three groups: A, C, D, E, K and P as the first group, B, G, O and Q as the second, and F, H, J and M as the third. Systems N and L were not used in this part of the experiment. The order of samples was randomised. The participants heard each sample only once, and at the beginning of each sub part, i.e. a new group of samples, a sample of natural speech was played as a reference point. As none of the participants were experts on synthetic speech, we suspected certain confusion over what we are really asking them to judge. We thus decided to familiarize the participants with synthetic speech and explain on examples the notions used in the experiment. Thus, before the experiment began, participants heard a set of samples different from the ones used in the actual experiment. The samples exemplified speech that we considered robotic but with no audible joins, robotic and with audible joins, non-robotic and smooth, and non-robotic, but with bad joins. Before the experiments started, each subject confirmed that they understood what was meant by “robotic” and “joins”. The experiment was conducted in a quiet room under the experimenter's supervision, and the participants heard the samples through headphones.

In the first part of the experiment the participants were asked to rate the sixteen samples on the scale from 1 to 7 on the basis of how robotic the voice sounds, where 1 means "very buzzy and robotic" and 7- "not robotic at all". They were instructed that they should not pay any attention to other factors, like bad joins, and even if the speech is of very good quality otherwise, smooth, with no audible joins, but the sound is buzzy, muffled and robotic, they should rate it accordingly.

In the second part of the experiment the participants were asked to rate the fourteen samples, again on the scale from 1 to 7, but this time focusing on joins, with 1

meaning "no audible joins, speech very smooth", and 7- "clear, audible joins that make the sentence unintelligible". They were again instructed to avoid paying attention to factors other than the one in question, that might affect their judgements. Twenty subjects participated in the experiment, of which exactly ten were female, and ten were male. The subjects formed a fairly heterogeneous group, including people of different ages, backgrounds and speaking different varieties of English. Most subjects were in their twenties, but the age span was from 17 to 63, with the average age of 28 years. Their educational backgrounds also varied, ranging from people who left school at 16 to PhDs. In terms of the variety of English spoken, we had one speaker of Canadian English, one of Irish English, one of Australian English, three of Southern British English, six of American English, and eight of Scottish English, which also ensured a good coverage of most standard varieties. However, we did not notice any correlations between the subjects' age, educational background, sex, or the variety of English spoken, and their answers.

3.2 Results

The data we collected is ordinal, so we decide to use medians in our analysis (Stevens 1946). The following table presents the results of the MOS test for the two acoustic features identified as represented by the dimensions on the MDS plane:

Figure 3. Median results of the perceptual experiment.

System	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	Q
Dimension 1	2.5	4	3	4	4	6	7	7	1	2	N/A	2	N/A	4	2	5
Dimension 2	3	4	3	4	4	5	6	4	3.5	3	7	4	4.5	3.5	2	5.5

Using SPSS we perform Spearman's rho correlations of the results of the experiment with the coordinates on the MDS plane. Results for Dimension 1, which we identified as bad joins, correlate well with Mean Opinion Scores (-0.690) and similarity scores (-0.604), with correlations significant at 0.5 and 0.1 level, respectively. Quite obviously, the results for this dimension correlate well with WER and the coordinates for Dimension 1, (0.719 and 0.768 respectively at 0.5 level). First, the hypothesis is supported in that bad joins usually affect intelligibility Secondly, the subjects ordered system in a similar fashion to that presented on the MDS configuration. There is no correlation with the coordinates for the other dimension, which is understandable, as we assumed in our analysis that the two features are not interdependent. There is, however, a significant correlation between scores for Dimension 1 and Dimension 2 in the experiment (0.792 significant at 0.5 level). This is unwanted, and most probably a result of the observed difficulty that participants had in isolating the features they were asked to rate.

Scores for Dimension 2, the perceived roboticness of synthesised speech, correlate well with MOS (-.893 at 0.5 level) and similarity to the original speaker (-0.890 at 0.5 level). However, there is also a significant correlation with WER (0.603 at 0.1 level), and coordinates for both dimensions (0.897 at 0.5 level for dimension 1-joins, and 0.620 at 0.5 level for dimension 2-roboticness), which is not something we expected nor desired in support of our hypothesis.

The results of the experiment seem to confirm our hypothesis in that the two dimensions do indeed correspond to the perceived quality of the speech being robotic and to bad, audible joins. The scores correlated reasonably well with the coordinates, meaning that the participants ordered the systems similarly to the original order along the axes. Although we report other correlations that were not expected, we attribute them to the unreliability of the data, and while indicative of certain problems surrounding the experiment, we do not treat them as disproving the hypothesis.

Chapter 4. Discussion

4.1 Validity of perceptual evaluation for speech synthesis

Organizing and conducting the perceptual experiment we experienced most of the things that inspired this study: the trouble that the human factor brings into any research. First, the experiment was costly, as without the £5 reward we would have only been able to test a few close friends. Secondly, even with the financial incentive, not many people found £5 enough to be bothered. Thirdly, those who did were very often only motivated by the money, and did not see the task as an important scientific contribution. They were very often distracted, and did not think much of the scores they were assigning to samples. As a result, the subjects that at the time seemed more focused and genuinely interested in helping the experimenter, rated the systems more accurately in terms of representing the original orders from the graph. Although there were no good or bad answers, we suspect that the correlation could be a lot stronger if all subjects had followed the task in a more conscientious manner.

Also, it soon became apparent that our participants had trouble isolating the feature we asked them to rate. Thus, systems that were problematic in a sense that one feature was really good while the other really bad, received ratings poorer than expected on their good feature. Although care was taken to ensure that subjects understand the notions used in the experiment, like "bad joins", our impression is that rather often the participants did not quite understand what they were asked to do, and were too embarrassed to admit or ask additional questions, and decided to just go for it and hope for the best. This perhaps suggests that different techniques of experimental design be used. Nevertheless, there is quite a lot of variance, and scores at times range from 1 to 7 for the same system and dimension. In overall, however, we decided not to exclude any answers. The initial analysis was, after all, based also on subjective evaluation that was even less controllable and restrictive than ours. Also, there were no clear outliers. Rather, some subjects were more tuned to specific features and were

able to isolate and thus rate them more accurately, while others were influenced by other features. Then, some subjects tended to be more harsh than others. Some made more fine-grained distinctions between systems that fell mid-range, while others could not see much difference. We felt, however, that those were the artifacts of relying on people's subjective judgements rather than invalid opinions. And thus the above problems were true for not just one or two subjects, but appeared regularly in the data, we decided to report the results as they were, without excluding any ratings.

Unfortunately, the results of the experiment confirmed what was suspected at the time of conducting it- that the two dimensions will be difficult to evaluate in separation from one another. It is well documented that acoustic features are difficult to judge in isolation, and one characteristic can affect our judgements of another, e.g. Vainio et al (2002) reporting interdependence of prosody and the perceived segment quality in Finnish text-to-speech. This is a problem that is certainly difficult to overcome, and needs a very careful experimental design. Perhaps, in line with Christensen and Humes (1997), training the participants to attend to particular dimension under consideration would serve as a good answer to the problem. In their study (Christensen and Humes 1997), the possibility of training the listener to attend to to a particular acoustic feature was explored. Followed an MDS analysis of 27 stimuli, and identification of the dimensions, a three exemplar stimuli were created, that contained low, middle, and high values for each of the identified dimensions. The exemplar stimuli were assigned names: circle, triangle and square, respectively. Participants were then presented with a series of these stimuli (first a series of ten of each in succession, then a series of “circle, triangle, square” groups), while their assigned names were presented on a computer screen. After the subjects were familiarized with the stimuli, they were tested if they could assign correct names to them, when played 20 times each in a random order. When 90% accuracy was achieved, the participants progressed to assigning names to the original stimuli, played 80 times each. Fifteen subjects achieved such accuracy after one training session, another three subjects required another training session to achieve a similar

accuracy. Judging the original stimuli, it was checked at regular intervals if the 90% accuracy for exemplar stimuli was maintained, and it was.

Such approach offers one solution to the problem at the cost of creating another. Such training again makes the evaluation costly and time consuming.

4.2 Conclusion

In the above sections we examined different techniques currently used in speech synthesis evaluation, while exploring one in depth, basing our analysis on the data from a large scale speech synthesis evaluation challenge. We performed a Multidimensional Scaling analysis of the data, and identified two dimensions along which the systems participating in the evaluation differ in terms of the perceived naturalness of synthesised speech. The proposed analysis reflects well the technology used by each system, as well as being rather powerful when the data is explored visually along with a further auditory analysis. It proposes that people make judgements on the perceived naturalness on the basis of segmental quality (bad joins) and the perceived roboticness of the speech. This is also in line with the finding by Mayo et al (2005), who suggested that people make judgements along at least two dimensions- segmental quality and suprasegmental quality.

The perceptual experiment conducted to support or disprove our hypothesis supported those findings to an extent. Although the results were not ideal for the point we were trying to prove, we strongly believe that it is the result of less than perfect experimental design and the difficulty of the task that the participants had to perform, rather than flaws in the analysis. Still, our study shows that first, Multidimensional Scaling is an interesting technique in modelling human perceptual behaviour, and secondly, that evaluation of speech synthesis requires a lot of attention and research if we want to arrive at truly reliable methods. With initiatives like The Blizzard Challenge progress is definitely being made towards a better understanding of the issues in speech synthesis evaluation. We hope that future editions will attract

attention of even more research groups, and the lessons learned from the challenge will push the whole field forward and trigger further improvements in the technology.

References

Benoit, Daniel – Martine Grice – Valerie Hanzan. 1996. The SUS: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. In *Speech Communication 18* (1996), 381-392.

Bennett, C.L., 2005. "Large scale evaluation of corpus-based synthesizers: results and lessons from the Blizzard challenge 2005." In *Proceedings of the Interspeech 2005*, Lisbon, pp. 105-108.

Black, Alan W. and Keiichi Tokuda. 2005. *The Blizzard Challenge- 2005: evaluating corpus-based speech synthesis on common datasets.* In INTERSPEECH 2005, 77-80.

Black, Alan W., Keiichi Tokuda, Simon King, Michael Picheny, and Shinsuke Sakai. 2006. *The Blizzard Challenge 2007. Evaluating corpus-based speech synthesis on common databases.* http://www.festvox.org/blizzard/

Borg, Ingwer and Patrick J F Groenen. 2005. *Modern Multidimensional Scaling. Theory and applications.* New York: Springer.

Christensen, L. A. and L. E. Humes. 1997. "Identification of multidimensional stimuli containing speech cues and the effects of training." In *The Journal of the Acoustical Society of America,* October 1997, Vol. 102(4), 2297-2310

Clark, Robert A. J. and Kurt E. Dusterhoff. 1999. "Objective methods for evaluating synthetic intonation." In *Proceedings Eurospeech 1999* (4), 1623-1626.

Clark, Robert A. J., Korin Richmond, and Simon King. 2007 "Multisyn: Open-

domain unit selection for the Festival speech synthesis system." In *Speech Communication*, 49(4):317-330.

Edgington, M. 1997. “Investigating the limitations of concatenative systems.” In *Proceedings of Eurospeech'97*, Rhodes, Greece. September 1997.

House, Arthur S, Carl Williams, Michael H. L. Hecker and Karl D. Kryter. “Psychoacoustic Speech Tests: A Modified Rhyme Test Rhyme Test,” Tech. Doc. Rept. ESD-TDR-63-403, U.S. Air Force Systems Command, Hanscom Field, Electronics Systems Division, 1963.

Mayo, Catherine, Robert A. J. Clark and Simon King. 2005. "Multidimensional scaling of listener responses." In Proceedings Interspeech 2005, Lisbon, Portugal, September 2005.

Stevens, S. S. 1946. "On the theory of scales of measurement." In *Science*, Vol. 103, 677-680.

Toda, Tomoki, Hisashi Kawai, Minoru Tsuzaki, and Kiyohiro Shikano. 2002. “Perceptual evaluation of cost for segment selection in concatenative speech synthesis” In P*roceedings of 2002 IEEE Workshop on Speech Synthesis.*

Vainio, Martti, Juhani Järvikivi, Stefan Werner, Nicholas Volk, and Jarmo Välikangas. 2002. “Effect of Prosodic Naturalness on Segmental Acceptability in Synthetic Speech.” In *Proceedings of the 2002 IEEE Workshop on Speech Synthesis*, Santa Monica, California, September 2002.