Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 311–326
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

# Listeners' weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis

Catherine Mayo *, Robert A.J. Clark, Simon King

*Centre for Speech Technology Research, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh EH8 9AB, United Kingdom*

Received 15 April 2010; received in revised form 3 August 2010; accepted 6 October 2010
Available online 19 October 2010

**Abstract**

The quality of current commercial speech synthesis systems is now so high that system improvements are being made at subtle sub- and supra-segmental levels. Human perceptual evaluation of such subtle improvements requires a highly sophisticated level of perceptual attention to specific acoustic characteristics or cues. However, it is not well understood what acoustic cues listeners attend to by default when asked to evaluate synthetic speech. It may, therefore, be potentially quite difficult to design an evaluation method that allows listeners to concentrate on only one dimension of the signal, while ignoring others that are perceptually more important to them.

The aim of the current study was to determine which acoustic characteristics of unit-selection synthetic speech are most salient to listeners when evaluating the naturalness of such speech. This study made use of multidimensional scaling techniques to analyse listeners' pairwise comparisons of synthetic speech sentences. Results indicate that listeners place a great deal of perceptual importance on the presence of artifacts and discontinuities in the speech, somewhat less importance on aspects of segmental quality, and very little importance on stress/intonation appropriateness. These relative differences in importance will impact on listeners' ability to attend to these different acoustic characteristics of synthetic speech, and should therefore be taken into account when designing appropriate methods of synthetic speech evaluation.


*Keywords:* Speech synthesis; Evaluation; Speech perception; Acoustic cue weighting; Multidimensional scaling

## 1. Introduction

Evaluation of the quality of output produced by a speech synthesis system is an important part of the design of a successful system. At the most basic level, evaluation can give system designers feedback on whether changes to a system engender an overall improvement in perceived quality of the output. At a more sophisticated level, evaluation could identify areas for further improvement. There are currently two main approaches to evaluating the quality of synthetic speech: (i) subjective, or human perceptual, methods, in which participants listen to examples of the synthetic speech and make judgements about quality (usually based on specific criteria) and (ii) objective, or computational, methods, in which models are built to automatically assess improvements to the synthesis system.

There are drawbacks to both types of analysis. Subjective perceptual evaluation requires the participation of numerous listeners in order to achieve statistical validity, and can thus be costly and time-consuming. In addition, listeners do not always achieve high levels of agreement either with each other or with themselves (Kreiman and Gerratt, 2000; Kreiman et al., 2007). This lack of reliability can make it difficult to draw meaningful conclusions from the results of subjective evaluation studies.

However, objective evaluation of synthetic speech is also problematic. In fields that make heavy use of objective

* Corresponding author. Tel.: +44 (0) 131 650 4434/651 1767; fax: +44 (0) 131 650 6626.
*E-mail addresses:* catherin@ling.ed.ac.uk (C. Mayo), robert@cstr.ed.ac.uk (R.A.J. Clark), simon.king@ed.ac.uk (S. King).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.10.003

evaluation measures (such as speech recognition), there is generally a non-opinion-based target with which to compare the output of the system—for example, the output of a speech recogniser can be compared against the text of the input given to the recogniser. Success in such fields is judged based on how well the output matches the desired target—in the case of a recogniser, that is typically how many words the recogniser correctly identifies. Judging the perceived quality of a speech synthesis system, on the other hand, is less straightforward. It *is* possible to make a direct comparison between acoustic characteristics of a target utterance (what the synthesiser has been asked to produce) and acoustic characteristics of the same utterance spoken by a human speaker (e.g. Clark and Dusterhoff, 1999). However, this ignores the fact that speech is highly variable (utterance-to-utterance, speaker-to-speaker, etc.) and that there are often many acceptable ways of producing a single utterance (see e.g. Jusczyk, 1997). As a result, it is possible for listeners to judge two utterances that are acoustically very different as being the same in terms of quality. The perceived quality of a synthetic speech utterance is clearly not, therefore, simply a matter of the degree to which the physical characteristics of the utterance match the physical characteristics of one single natural speech utterance. Rather, the perceived quality of an utterance is a *psycho*-physical construct, which is closely tied to both the physical, acoustic characteristics of the utterance being judged, *and* to listeners' psychological responses to these characteristics (Kreiman and Gerratt, 1998). Thus the success of an objective measure of synthetic speech quality depends on two things: (i) how well the measure models the physical characteristics of the speech being evaluated and (ii) how well the measure models listeners' behaviour with respect to that speech.

There have been a number of attempts to model human perceptual evaluation of speech. However, to date, only low to moderate correlations have been found between objective and subjective ratings of speech quality for both synthetic speech (Chen and Campbell, 1999; Clark and Dusterhoff, 1999; Falk et al., 2008; Klabbers and Veldhuis, 1998; Stylianou and Syrdal, 2001; Vepa and King, 2004; Wouters and Macon, 1998) and for natural speech (Rabinov et al., 1995). Some higher levels of correlation have been found, but only for very restricted speech sets (e.g., isolated words rather than full sentences (Cerňak and Rusko, 2005; Cerňak et al., 2009)). This lack of correlation seems to stem not from an inability to model the physical characteristics of speech, but from difficulties in modelling human perceptual responses. As noted above, listeners' subjective evaluations often lack strong inter- and intra-rater consistency. In addition to making interpretation of subjective evaluations difficult, such inconsistent behaviour is inherently less amenable to objective computational modelling.

The question to be answered, therefore, is why subjective evaluation behaviour is so inconsistent. There is an understanding in the field of speech synthesis of what paradigms are currently available for testing perceived quality (e.g. Expert Advisory Group on Language Engineering Standards, 1996), and an understanding of the need for principled use of evaluation paradigms (e.g. Bailly et al., 2003). This knowledge should allow for studies to be carried out in which listeners' responses are more consistent. However, despite the awareness of testing paradigms, there is a general lack of understanding of the psycho-acoustic processes that underpin the complex task of auditory evaluation of synthetic speech. In particular, it is not clear from research to date what the exact relationship is between the acoustic characteristics of synthetic speech and listener responses to these characteristics. Without an understanding of this relationship, it is very difficult to choose evaluation methods in a principled manner, and as a result, raters may be asked to carry out tasks which their perceptual systems cannot physically perform: naturally this could easily result in inconsistent or unexpected rating behaviour.

In fact, what little is known about the psycho-acoustic task of speech evaluation does point to the possibility that listeners are often asked to perform perceptually challenging tasks. One of the dominant state-of-the-art speech synthesis techniques (and the one which we use in this work) involves *unit selection*. In this method, a large database of natural speech is labelled in terms of units (usually phones, diphones or half phones). An automatic search is then performed to find the best units or sequences of units from the database, and these units are concatenated together to form the target utterance (see e.g. Black et al., 1997–2004). This method has resolved many of the issues surrounding gross segmental quality and intelligibility that caused problems for earlier, rule-based synthesis systems, and has thus allowed researchers to move on to fine-tuning individual sub-segmental characteristics (e.g., discontinuities at concatenation points, Klabbers and Veldhuis, 2001) or supra-segmental characteristics (e.g., intonation, Clark, 2003). As a result, speech synthesis evaluation is now much less about determining the overall intelligibility or overall acceptability of synthetic speech, and more about evaluating the quality of a single one of these sub- and supra-segmental characteristics. Unfortunately, research has shown that listeners sometimes find it difficult to focus on just one characteristic, particularly when faced with complex acoustic stimuli such as speech. For example, it has been found that listeners are much less able to rate intonation consistently when it varies simultaneously with many other acoustic characteristics than when intonation is the only acoustic characteristic of the stimulus set to be varied (Kreiman and Gerratt, 2000). This would suggest that it may be beyond listeners' abilities to evaluate just one sub- or supra-segmental characteristic of a synthetic speech utterance.

However, concluding that subjective evaluation of single characteristics of multidimensional stimuli is impossible assumes that listeners give equal perceptual attention or

"weight"[1] to all acoustic characteristics and cannot disentangle individual characteristics from each other. This does not seem to be the case. Instead, it appears that listeners give more weight to some characteristics than others. Evidence for this comes from studies that have shown, for example, that the addition of synthetic intonation to natural speech segments was more detrimental to listeners' speech quality ratings than was the addition of synthetic segment duration (Plumpe and Meredith, 1998), suggesting that for this listening situation, intonation was weighted more heavily than segment duration. Other studies have found that listeners have been more influenced by intonation appropriateness than by segmental quality (Vainio et al., 2002), less influenced by intonation naturalness than by segmental quality (Hirst et al., 1998; Vainio et al., 2002), and less influenced by intonation variation than by overall synthesis quality or presence of discontinuities due to concatenation (Jilka et al., 2003; Syrdal and Jilka, 2004). In fact, studies from across perceptual evaluation, speech perception and general auditory perception research (e.g. Allen and Scollie, 2002; Best et al., 1981; Christensen and Humes, 1997; Mayo and Turk, 2004; Nittrouer, 2004; Syrdal and Jilka, 2004) indicate that listeners have very complex hierarchies of weighting for different acoustic characteristics, and that these can differ depending on the stimuli (e.g., speech versus non-speech, natural speech versus synthetic speech, first language versus second language, etc.). This would suggest two things. First, it should be straightforward to devise an evaluation paradigm to assess the quality of those acoustic characteristics that are by default weighted heavily by listeners. Second, if listeners are asked to evaluate characteristics that are lower on their weighting hierarchy, they could easily be unduly influenced by those characteristics that are weighted more heavily. This could leave a large number of acoustic characteristics unaccessible for evaluation.

Fortunately, research suggests that even those acoustic characteristics that are weighted less heavily could be accessible to listeners, under specific circumstances. Studies have found that listeners' auditory weighting patterns are not necessarily fixed, but rather can be flexible, with most adult listeners changing their weighting hierarchy to meet the needs of the stimuli or the listening situation. A number of studies have examined the ease with which listeners can detect spectral discontinuities at join points (points at which units of speech from two different source utterances are concatenated). These studies have found differences in rates of detection for male and female synthetic voices, and for joins in different segmental contexts (e.g., monophthong versus diphthong, pre-vocalic consonants versus post-vocalic consonants, front versus back vowels, etc., see Klabbers and Veldhuis, 2001; Syrdal, 2001). Additionally, listeners' default weighting patterns for speech and non-speech auditory stimuli have been shown to be capable of change due to outside influences: listeners' acoustic characteristic hierarchies have been changed by training (Christensen and Humes, 1997; Francis et al., 2008; Iverson et al., 2005), by manipulation of the stimuli to mask certain dimensions (e.g., simultaneous white noise, Wardrip-Fruin, 1982; Wardrip-Fruin, 1985, or reverberation, Watson, 1997) or to enhance certain dimensions (Hazan et al., 1998; Hazan and Simpson, 1998), by manipulation of the distribution of the acoustic dimensions across the whole stimulus set (Allen and Scollie, 2002) or by manipulation of listeners' conscious focus of attention (e.g., by presenting the rating task simultaneously with another task (Gordon et al., 1993)). It appears possible, therefore, that if listeners do not by default give adequate attention to the acoustic dimension under investigation, appropriate methods could be designed to allow listeners to refocus their attention.

However, both (i) taking advantage of listeners' default hierarchy of weighting to evaluate dimensions that are weighted heavily and (ii) developing methods that will allow listeners to attend more closely to dimensions that receive less weight, presuppose a better understanding of the default weighting given to potential acoustic dimensions of synthetic speech than is currently available. The study described here is an attempt to provide a more complete account of the acoustic dimension weighting behaviour of listeners with respect to unit-selection synthetic speech.

### 1.1. The current study

The current study aimed to determine which of a large number of acoustic characteristics of synthetic speech play greater or lessor roles in listeners' responses to such speech. To address this aim, the study made use of multidimensional scaling (MDS) techniques (Kruskal and Wish, 1978), which have been used successfully to determine weighting patterns for numerous auditory domains: e.g., complex non-speech sounds (Allen and Bond, 1997; Allen and Scollie, 2002), coded speech (Hall, 2001), segmental contrasts (Iverson et al., 2002), voice quality (Kreiman and Gerratt, 2004), and musical timbre (Marozeau et al., 2003).

The goal of MDS techniques is to create a geometric, spatial representation of the proximities between objects—either actual or perceived proximities. In the current study, synthetic utterances were presented pairwise to listeners, who were asked to decide whether each pair was "similar" or "different" in terms of naturalness. The total number of times a given pair of utterances received a "different" response from listeners was taken as the proximity value for (i.e., the psycho-physical distance between) that pair of utterances.

[1] The term perceptual weight is used here, and throughout this paper, as a synonym for perceptual attention or importance, and is used in a relative, rather than absolute or quantifiable sense (see Francis et al., 2008; Hazan and Barrett, 2000; Iverson et al., 2005; Mayo and Turk, 2004; Mayo and Turk, 2005).

The output of an MDS analysis is an *n*-dimensional stimulus space or map, in which each object—here, each of the synthetic utterances—is represented by a single point. The first key characteristic of an MDS map is that two objects that are physically or psycho-physically close are represented by two points that are close on the map, while two objects that are physically or psycho-physically distant are represented by two points that are farther apart on the MDS map. The second important characteristic of an MDS map is that the dimensions that make up the space often correspond (directly or indirectly) to the physical or psycho-physical dimensions used most heavily by subjects to make their proximity judgements. Thus by examining and interpreting both the MDS map dimensions and the configuration of the points within those dimensions, it should be possible to determine the underlying characteristics of the objects represented in the space that led to subjects' responses to those objects. For the current study, this means that analysis of the MDS space should allow us to identify some of the acoustic characteristics of synthetic speech that relate to listeners' judgements of that speech. Analysis and interpretation of an MDS stimulus map can be done in a number of ways. For many two- and potentially three-dimensional spaces it may be possible to interpret the space by visually examining the distribution of the objects within the space and trying to find any underlying pattern(s) in the organisation of the points. For stimulus spaces with four or more dimensions (i.e., numbers of dimensions that are less straightforward to represent in a way that can be examined visually) or for more complex two- and three-dimensional maps, it is often necessary to make use of other methods, such as cluster analysis, or multiple regression (i.e., regressing relevant characteristics of the objects in the MDS space on the coordinates of the points in the MDS map). In the current study, it was not clear before MDS analysis was carried out whether the MDS map resulting from listeners' "different" responses would be straightforward to analyse either visually or auditorily. Therefore, in addition to the listening experiment, a thorough acoustic analysis of the synthetic utterances used in the listening experiment was also carried out, with the aim of identifying and quantifying a large number of acoustic characteristics of the synthetic speech with which to compare the MDS stimulus space.

A small pilot study using MDS methods (Mayo et al., 2005), indicated that listeners are influenced by at least two general characteristics when determining how similar utterances sound in terms of naturalness: segmental information (likely to include such things as appropriateness of segments, audibility of joins at unit edges, and number of discrete units used to create the utterance), and prosodic information (most likely including intonation, duration/timing, and stress). The small number of utterances used in this pilot meant that stimuli tended to be placed at extremes in the stimulus space, which in turn meant that it was not possible to determine whether listeners might respond to more fine-grained aspects of the two general characteristics identified. However, despite this, the study showed clearly that MDS techniques are likely to serve as a useful method of identifying the acoustic characteristics of synthetic speech that most influence listeners' perceptual evaluation behaviour. The current study aimed to expand on the pilot by increasing the number of utterances and the number of listeners in order to engender a more fine-grained stimulus space, and thus identify more precise characteristics which might be involved in the perception of naturalness in synthetic speech. Note that the experiment described in the current study focused on listeners' evaluation of unit-selection speech produced by a typical unit selection speech synthesis system, Festival (Clark et al., 2007); further studies should endeavour to examine listeners' responses to other unit-selection synthesis systems, as well as systems based on other techniques (e.g. statistical parametric models, Zen et al., 2007).

In summary, the aim of the current study was to better understand the relationship between the acoustic characteristics of unit-selection synthetic speech and listeners' responses to such speech. In the current study, 24 synthetic utterances were presented pairwise to listeners for assessment as either "similar" or "different" in naturalness. Listeners' responses were analysed using MDS techniques, and the resulting spatial map was correlated with acoustic characteristics of the 24 synthetic utterances. These methods should allow us to determine which of the acoustic characteristics identified in the synthetic speech played a role in listeners' perception of synthetic speech quality, and should also allow us to begin to evaluate these characteristics' relative importance in determining how similar/dissimilar listeners perceived the stimuli to be.

## 2. Method

### 2.1. Participants

Thirty adults (22 female, 8 male[2]) ranging in age from 18 years to 43 years (average age 24 years) took part in this experiment. All were monolingual native speakers of English who were living in Edinburgh, UK at the time of testing (self-reported dialects broke down as follows: Southern British English – 12; Northern English – 6; Scottish – 5; Irish – 3; North American – 4). All subjects reported themselves and their siblings as being free from speech/language disorders and dyslexia, and all reported being free from hearing deficits. Immediately prior to testing, each participant was required to pass a hearing screening at $\leqslant$35 dB HL in the range 125 Hz–8 kHz. All listeners were paid for their time.

[2] This gender split accurately reflects the gender split of the undergraduate students in the department from which the majority of the subjects were solicited.

### 2.2. Stimuli

Stimuli were generated using the multisyn module of the Festival speech synthesis system, which is a typical unit-selection text-to-speech synthesiser. Unit-selection synthesis uses a large database of natural speech recordings of a single speaker, from which suitable units are selected and concatenated to create any required target utterance. In Festival, the units are diphones: units of speech that begin halfway through one phone and finish halfway through the subsequent phone. The speech database is automatically phonetically labelled and from the phone labels, diphone boundaries are derived.

At synthesis time, Festival first performs text analysis, resulting in a phonetic transcription of the utterance to be created; this transcription is then converted to a sequence of diphones known as 'targets'. Then the best sequence of diphone instances from the database is selected and concatenated to create the output speech. As in other unit-selection synthesisers, the measure of 'best' has two components. One, the *target cost*, measures the degree of mismatch between each candidate diphone available in the database and the corresponding target diphone. The mismatch is judged in terms of linguistic context—that is, the difference between the phonetic and prosodic contexts of candidate and target—on the assumption that fewer mismatches will result in perceptually more natural speech output. The other component is the *join cost*, which measures how well each pair of consecutive candidates will concatenate; join cost is typically calculated based on properties of the speech signal such as spectral discontinuity, energy and F0. It is assumed that smaller differences in these acoustic properties at join points between candidate units will be less perceptually noticeable and thus more natural-sounding. A search must be performed in order to find, from amongst all possible sequences of candidates, the one that minimises the sum of target and join costs. By defining the join cost between candidates that are consecutive in the database to be zero, the search is biased towards selecting contiguous stretches of speech. This means that there will generally be fewer concatenation points in the output speech than there are phonemes in the target utterance. Festival additionally attempts to avoid using extreme outlier candidates (in terms of duration, for example) and can substitute any missing diphones with alternatives (i.e., in cases where there are no instances of a required diphone type in the database).

The text material for the synthetic utterances used in this study came from the "phonetically-compact" portion of the TIMIT corpus (Garofolo, 1988). This portion was hand-designed "to include a basic phonetic coverage and interesting phonetic environments" with special regard for "phoneme pair coverage, consonant sequence coverage and the potential for applying phonological rules both within words and across word boundaries" (Lamel et al., 1989, p. 2162). The reason for using this material was to ensure the test utterances would contain a reasonably wide variety of acoustic–phonetic phenomena that might impact on perceived speech quality.

All 450 of the phonetically-compact TIMIT sentences were synthesised using the Festival 1.96 multisyn engine with a female, Standard Southern English voice (cstr_rpx_nina_multisyn).[3] Twenty-four of the resulting utterances were chosen as the test utterances for the current study. All 24 were chosen at random, with the proviso that they be similar in duration (both in milliseconds and in total syllables; this was done in order to minimise any effect of length of utterance): the test utterances were therefore between 2.60 and 2.70 s in duration (average duration: 2.65 s) and all contained approximately 12 syllables (range: 9–14 syllables; average: 12 syllables). None of the 24 test utterances were manipulated during or after synthesis. All variations in the acoustic characteristics of the synthetic speech and in the perceived naturalness of that speech were thus uncontrolled, and resulted only from the diphone coverage of the voice database used and the performance of Festival.

In addition to the 24 test utterances, a set of 10 practice utterances and a set of six familiarisation utterances (both also from the phonetically-compact TIMIT sentences) were also synthesised.

The 10 practice utterances were between 2.26 and 2.27 s in duration (average duration: 2.26 s) and all contained approximately 10 syllables (range: 6–13 syllables; average: 10 syllables). Again, none of the practice utterances were manipulated during or after synthesis. All of the sentences were lexically different from those presented in the main test.

The six familiarisation utterances were designed to be presented in three specific pairs to illustrate extreme examples of "similar" and "different" with regard to naturalness. These pairs comprised one pair of "different" utterances, which contained an extremely natural utterance and an extremely unnatural utterance, and two pairs of "similar" utterances, one of which contained two extremely natural utterances, and the other of which contained two extremely unnatural utterances. The six utterances that made up these familiarisation pairs were synthesised using the same voice that was used to create the practice utterances and the main test utterances. In order to create the three utterances designed to illustrate extremely unnatural speech, only a subset of the voice database was employed by the synthesiser (400 utterances, rather than the full database of 2000 utterances). Restricting the diphone coverage available to the synthesiser degrades the output speech quality. The three utterances designed to illustrate very natural synthetic speech were chosen by the experimenters as being representative of the most natural utterances that this synthesis system was capable of producing from the given voice database. None of the familiarisation utterances were lexically the same as those in either the practice or the main test.

[3] A live demonstration of this voice, called 'Nina', can be found at http://www.cstr.ed.ac.uk/projects/festival/onlinedemo.html.

#### 2.2.1. Test sets

In order to acquire the proximity values necessary for MDS analysis, it was necessary to present the synthetic speech utterances in pairs, to determine the perceived degree of similarity/dissimilarity in naturalness between each utterance and every other utterance. To create the utterance pairs for presentation to the listeners in this study, each of the 24 test utterances was paired with every other test utterance, in both directions (i.e., both AB and BA), resulting in 576 possible pairs of utterances (24 of which were same-utterance pairs, e.g., utterance 1 paired with utterance 1). Pilot testing (Mayo et al., 2005) had revealed that listeners could rate 168 pairs of utterances in approximately 30–40 min, and that this number of utterance pairs and duration of testing was approaching the upper limit of what listeners could tolerate in a single test session. Therefore, to avoid listener fatigue and any possible associated drop in performance, it was decided that the complete set of 576 utterance pairs would be divided into three smaller test sets, and that each listener would hear only one of these three test sets.[4]

The subdivision of the complete set of utterance pairs was done as follows. A $24 \times 24$ grid of all possible pairs of utterances was created. This complete set of pairs was then subdivided into nine smaller subsets by superimposing a $3 \times 3$ grid over the larger $24 \times 24$ grid. Each of the nine subsets therefore contained 64 pairs of utterances (e.g., utterances 1–8 $\times$ utterances 1–8, or utterances 1–8 $\times$ utterances 9–16, or utterances 1–8 $\times$ utterances 17–24, etc.). A Latin square arrangement was then used to allocate the utterance pairs in the nine subsets into one of three test sets: the cells in the $3 \times 3$ grid were labelled A–B–C, B–C–A, and C–A–B by row, and all of the utterance pairs in the cells labelled 'A' were put into a single test set; the same was done for the utterance pairs in the cells labelled 'B' and those in the cells labelled 'C'. This resulted in 192 utterance pairs per test set. At this point all of the same-utterance pairs were removed from the test sets (there were eight same-utterance pairs per test set). This was done because all utterance pairs *except* the same utterance pairs contained two lexically different sentences. Including a small number of pairs of lexically identical sentences in a test set made up predominantly of different-utterance pairs could have biased listeners' responses in an unpredictable way. The final version of each of the three test sets therefore contained 184 pairs of utterances.

This design resulted in a manageable test set size for each listener, while maintaining maximal variation in what each listener heard (each test set contained each of the 24 test utterances paired with one third of the 24 test utterances, less the same-utterance pairs), and ensuring complete coverage of the test set (i.e., ensuring that every possible different utterance pair was heard by at least a portion of the subjects).

[4] A possible alternative solution, that of presenting all utterance pairs to every listener over multiple test sessions, was dismissed due to the high probability of subject attrition over second and subsequent test sessions.

### 2.3. Presentation of stimuli

All participants were tested individually in a sound-treated room. The stimuli were presented over closed-back headphones (Sennheiser Evolution H270, frequency response 12–22,000 Hz) with volume set at a constant, comfortable listening level. Testing took place in one 40–45 min session, with two short breaks part way through testing.

Presentation of the pairs of stimuli was controlled by a suite of computer software (E-Prime, Schnieder et al., 2002) that also recorded listeners' responses. The listeners' task was to listen to the two utterances in a pair, and indicate whether they were "similar" or "different" in terms of their naturalness (that is, with regard to how much like real speech each utterance seemed to be). Importantly, the participants were not instructed to listen to any one acoustic characteristic of the stimuli, or to any specific psychoacoustic concept that might make up the overall construct of "naturalness" (e.g., "listening effort," "pleasantness," "pronunciation," etc.) such as used in mean opinion scale (MOS) tests of synthetic speech quality (ITU-T Recommendation P.85, 1994). Instead, the listeners' task was simply to make a binary decision ("similar" or "different") about the degree of similarity in naturalness of each pair of stimuli.

Before testing, the participants were given an opportunity to listen to the three familiarisation pairs of synthetic sentences. As noted in Section 2.2, these pairs were designed to introduce listeners to extreme examples of "similarity" and "differentness" in naturalness, and were explicitly identified to the listeners as such (that is, listeners were told "These two sentences are similar in terms of naturalness."). No responses were collected from listeners during this period. Following this familiarisation, a pre-test was administered, to ensure the listeners understood the task. This pre-test consisted of the 10 practice utterances, again presented in pairs. Only 10 of the possible pairs of utterances were presented to each listener, in random order. No feedback regarding naturalness was given during this pre-test; additionally, as there was no pre-defined "correct" response to whether a given pair of utterances sounded "similar" or "different", no listeners were excluded on the basis of answers given during this pre-test period.

During the test proper, each participant heard one test set of 184 pairs of utterances (see Section 2.2.1 for a description of these test sets). The interval between the presentation of each member of a pair was 5000 ms (onset to onset). Responses were not timed, and the presentation of the next pair of utterances began 2000 ms following the entry of the preceding response. Breaks were given following the presentation of the 61st and the 122nd pairs; the duration of these breaks was controlled by the participants.

### 2.4. Acoustic analysis of stimuli

It was unclear at the start of the current study which potential acoustic characteristics might play a role in listeners' judgements of the stimuli used in this study. It was therefore important to undertake as complete an analysis as possible of the stimuli to determine the spread of acoustic characteristics across the utterances and subsequently to use this to interpret the MDS configuration.

Two types of acoustic analysis were carried out—automatic and manual—on the 24 synthetic utterances used in the listening experiment, and where appropriate, on the natural utterances used to create the synthetic utterances. All analyses are listed in Appendix A. The automatic analyses included measurements carried out by Festival in the process of synthesising the 24 stimulus utterances, which included target cost of individual units, join cost for joining two individual units, and number of "bad" units (those units that are not statistically representative of other units containing the same segments from the database) or missing units (those diphones that cannot be found in the database and are replaced by something similar, e.g., a diphone containing a long vowel may get backed-off to a unlengthened version).

Also in the automatic analysis were those measures that could be automatically calculated based on Festival's measurements. These were: total cost (a measure based on overall target cost and overall join cost for an utterance); target cost for different types of target diphone (e.g., consonant–consonant, vowel–consonant, etc.); join cost for different types of join (e.g., joins in vowels, stop consonants, fricatives, etc.); proportion of units or joins with different values of target or join cost. Finally, the automatic analysis included a number of other automatically calculable measurements, for example: duration of an utterance in milliseconds, whether the initial and final diphones were taken from initial/final position in the source utterance, etc.

In addition to these automatic analyses, five different manual analyses were also carried out, three on the 24 synthetic utterances, and two on each of the natural source speech utterances in the voice database from which units were selected to create the 24 synthetic utterances. The natural source utterances were first analysed for number of transcription/pronunciation errors per synthetic utterance. The database of natural speech utterances is automatically phonetically labelled, based on the text that the database speaker was given to read. If the speaker deviates from this text in some unexpected way, this will cause a mismatch between the database speech, and the phonetic label, which in turn will cause problems if these mismatched units are chosen for use in synthesis. Errors in this category included (i) mispronunciations by the database speaker (e.g., pronouncing "Maryland" as "Marilyn": results in a section of database labelled /d/ when no /d/ has been produced); (ii) automatic segment labelling not accounting for pronunciation variation (e.g., speaker producing "window" as /wɪndə/, but the lexical entry only containing an unreduced diphthong in the second syllable: results in errors in contexts where reduction of the diphthong is not possible, e.g., "gateau"); (iii) the incorrect labelling of any non-silent stretch of speech as "silence".

The natural source utterances were also manually analysed for number of segmentation errors per synthetic utterance. As noted above, Festival makes use of segmentation information to determine the points at which a unit should be cut out of a source utterance for use in a target utterance. Note that cuts/joins are not made at boundaries between phones, but at diphone boundaries. These points are, however, derived from phone segmentation information. The accuracy of the initial segmentation of the natural speech is therefore important to the success of the splicing procedure in unit selection speech synthesis. Accuracy of the segmentation of the natural speech database was evaluated using the principles set out by Turk et al. (2006). Segmentation errors were only noted if they occurred at the edge of a unit boundary (e.g., if the automatic segmentation was found to be incorrect for the boundary between /ɹ/ and /g/, but the entire sequence /aɹgə/ was selected from a source utterance for use in a target utterance, then this error in segmentation was ignored).

The synthetic utterances were manually analysed for: (i) number of inappropriate diphones, (ii) number of detectable joins, and (iii) number of areas of incorrect stress/intonation (both lexical and phrasal). Classed as inappropriate diphones were any diphones taken from a segmental context in the source utterance which would be segmentally incompatible with the target utterance (e.g., a post-vocalic /ɹɪ/ diphone taken from the word "prio*ri*tization" to be used in the post-stop context in the word "p*re*dicament": the post-vocalic context would engender a much longer /ɹ/ than would be appropriate for the post-stop context). Detectability of joins were analysed by means of spectrograms and spectra, and "detectable joins" included all joins in which there was at least one of: (i) a mismatch in vowel formant frequencies at the join point, (ii) a mismatch in pitch at the join point, (iii) a mismatch in intensity at the join point. Stress/intonation was marked as inappropriate in the following circumstances: (i) if the source utterance carried/did not carry lexical stress on the to-be-used vowel and this was inappropriate for the target context (e.g., the lexically stressed syllable /bɪɪ/ from the word "*BRI*tish," being used in the target word "UP*bri*nging," where the syllable should not carry lexical stress), (ii) if the source contained stress/intonation appropriate for a particular location in an utterance and this was inappropriate for the location in the target utterance (e.g., a sequence with phrase-final falling intonation used phrase medially) and (iii) if the source carried/did not carry contrastive stress and this was inappropriate for the target context.

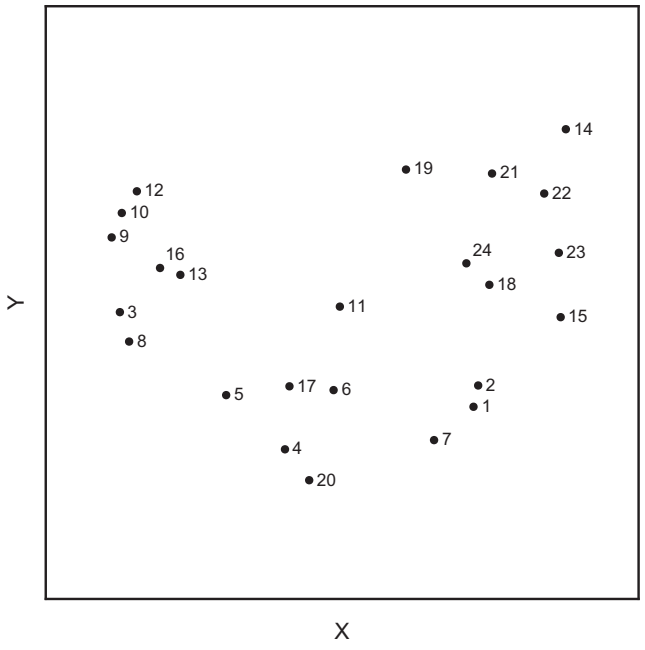


Fig. 1. Two-dimensional $X$–$Y$ projection of the three-dimensional MDS map. Each number refers to the number of a single synthetic utterance.

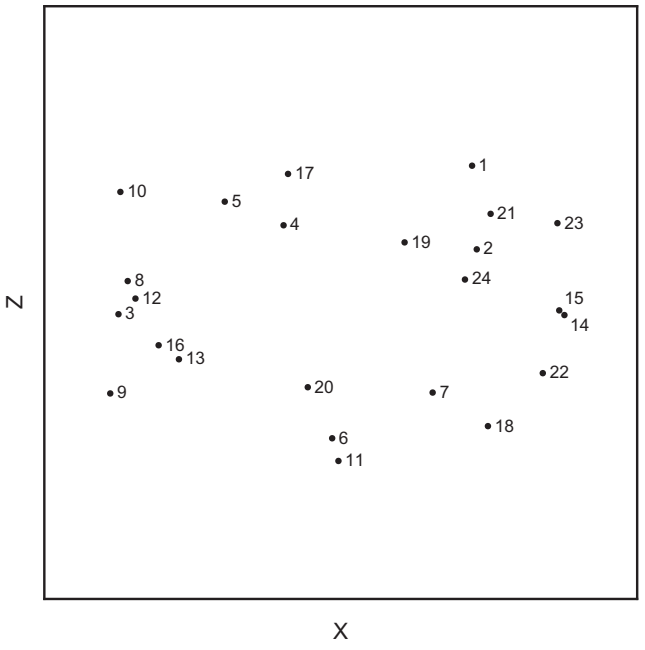


Fig. 2. Two-dimensional $X$–$Z$ projection of the three-dimensional MDS map. Each number refers to the number of a single synthetic utterance.

## 3. Results and discussion

The proximity values dictated by the listeners' "similar"/"different" responses to the 24 synthetic utterances used in this study are best represented by a three-dimensional MDS map [$R^2 = .75$, residual stress = .18]. As the rotation of the MDS space is potentially arbitrary, principal component analysis (PCA) was used to ensure any principal components were aligned with the axes; PCA did not further rotate the space.[5] Figs. 1–3 show three two-dimensional pairwise axis projections of the three-dimensional MDS map.

### 3.1. Visual, auditory and cluster analysis

Kruskal and Wish (1978) advocate attempting a basic visual analysis of MDS maps in the first instance, before subjecting the map to further analysis. In the case of the current study, this basic analysis also included an auditory analysis and a cluster analysis of the distribution of the synthetic utterances within the MDS space.

From an initial visual and listening analysis, it is clear that listeners are able to make a certain number of important distinctions with regard to the acoustic characteristics of the synthetic speech used in this study. We identified five general groups into which listeners appear to have classified the utterances. More fine-grained groupings may be present in this data, however they are not immediately present from this first-pass visual and auditory analysis: they may become evident after further (statistical) analysis (see below). These five groupings are outlined below. To illustrate each grouping, audio files containing one synthetic utterance from each group accompany the electronic version of this manuscript. To access these files, click on the link following the description of each group.

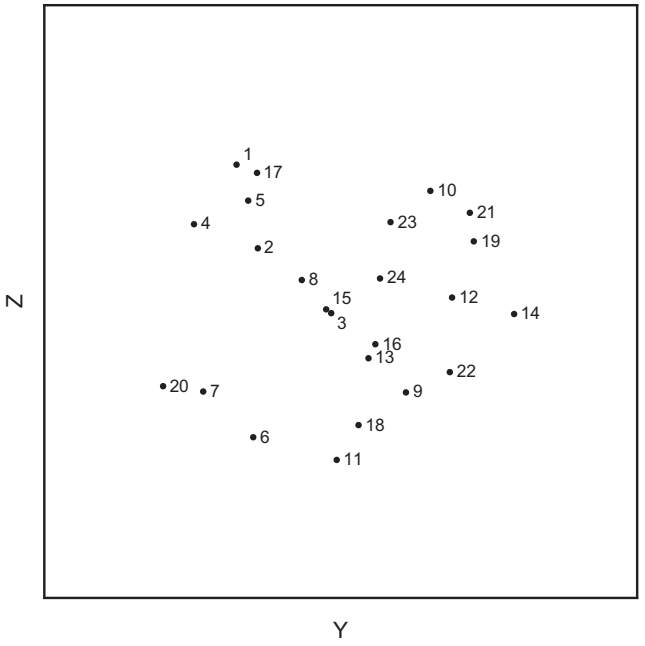


Fig. 3. Two-dimensional $Y$–$Z$ projection of the three-dimensional MDS map. Each number refers to the number of a single synthetic utterance.

[5] The fact that PCA did not rotate the MDS space suggests that the way that MDS is implemented in SPSS incorporates PCA, which is an entirely reasonable assumption.

(A) very high quality utterances, with (i) very few poor units (i.e., few units with transcription/pronunciation errors, inappropriate diphones, high target costs, etc.) and (ii) very few joins (i.e., the entire utterance was made up of long stretches of contiguous speech from the database), for example, utterances 9, 10, 12. Click on the following link to listen to utterance 10: "Steph could barely handle the psychological trauma." (see Supplementary Audio 1);

(B) utterances with very poor quality joins (as indicated by high join costs and audible discontinuities at joins), for example, utterances 15, 18, 21, 22, 23. Click on the following link to listen to utterance 15: "We saw eight tiny icicles below our roof." (see Supplementary Audio 2);

(C) utterances with many joins, but all of them of relatively high quality (indicated by low join costs and absence of audible discontinuities at joins), for example, utterances 6, 11, 20. Click on the following link to listen to utterance 6: "His scalp was blistered from today's hot sun." (see Supplementary Audio 3);

(D) utterances with a handful of poorly selected units (containing instances of transcription/pronunciation errors, etc.), for example, utterances 4, 5, 17. Click on the following link to listen to utterance 5: "Will you please describe the idiotic predicament." (see Supplementary Audio 4);

(E) utterances with many poorly selected units, for example utterances 1, 2. Click on the following link to listen to utterance 1: "It's fun to roast marshmallows on a gas burner." (see Supplementary Audio 5).

To attempt to validate these groupings as being meaningfully different in terms of their location in the MDS space we performed *k*-means clustering. The within-groups sum-of-squares for different cluster sizes suggests that four to six clusters are present. Clustering with six clusters produces groupings which match very closely our initial interpretation, as illustrated in Fig. 4. In the top left of Fig. 4 we see a cluster containing the stimuli 3, 8, 9, 10, 12, 13 and 16, which matches our group A. In the top right of Fig. 4 we see two clusters, the first containing stimuli 15, 18 and 22 and the second containing stimuli 14, 19, 21, 23 and 24. These two clusters together match our group B (note that these clusters merge, if clustering is performed with only four clusters). Stimuli 6, 7, 11 and 20 form a cluster at the bottom centre of Fig. 4, corresponding to our group C and stimuli 4, 5 and 17 form a cluster to the left of the previous cluster, relating to group D. Finally stimuli 1 and 2 form a cluster that matches the stimuli described by group E (again if only four clusters are requested, these last two clusters merge).

These groupings allow us to make the following inferences. First, listeners seem able to differentiate between utterances on the basis of the *number* of joins in an utterance (many, even if low cost/high quality, as in group C, versus few, as in group A) and the *quality* of joins across an utterance (poor quality/high cost, as in group B, versus better quality/lower cost, as in group C). Listeners also seem able to differentiate between utterances on the basis of number of poor quality units, with at least three different identifiable levels of unit quality: (i) many, poorly selected units, as in group E, (ii) a handful of poorly selected units, as in group D, (iii) few poor quality units, as in group A. It is not, however, evident from these general groupings whether listeners are able to, or do, use any aspect of intonation or prosody in their judgements of the naturalness of these utterances.

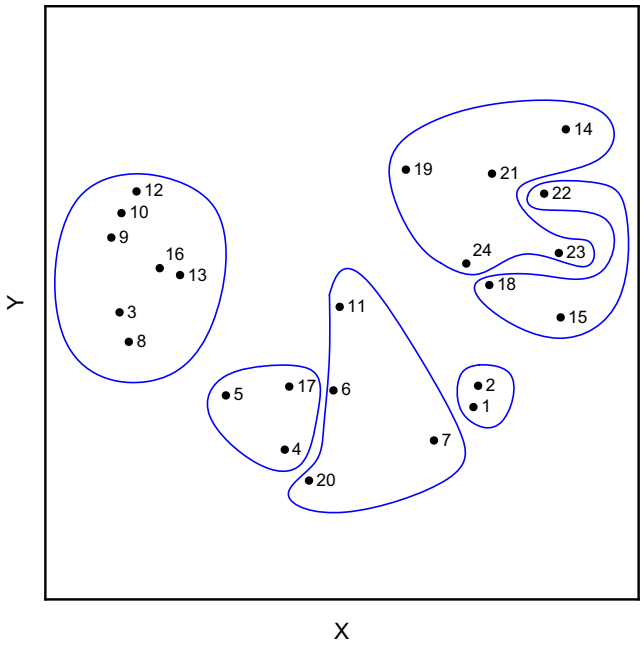


Fig. 4. *K*-means clustering of two-dimensional *X*–*Y* projection of the three-dimensional MDS map with six clusters. Each number refers to the number of a single synthetic utterance. Refer to the text for a description of the general acoustic characteristics of each cluster.

### 3.2. Regression analysis: accounting for the x-, y-, and z-dimensions

Following the above visual, auditory and clustering analyses, to further investigate the resulting spacial configuration, all of the acoustic characteristics of the synthetic speech utterances were regressed onto the *x*-, *y*- and *z*-dimensions of the MDS map. This analysis is performed to determine whether any meaningful interpretation of the axes can be made.

Note that while the configuration of utterances along both the *x*- and *y*-dimensions was found to correlate with a number of the acoustic characteristics at $p = 0.05$, the distribution of utterances along the *z*-dimension did not correlate with any characteristics at this level of *p*. At $p = 0.06$, the *z*-dimension was found to be accounted for by six acoustic characteristics; these results should be treated more cautiously, given the weaker relationship between the dimension and the characteristics. The results of all of these regressions can be found in Tables 1–3. $R^2$*change*

Table 1
Multiple regression analysis modelling MDS coordinate values from acoustic characteristics: *x*-dimension. See Appendix A for a description of each acoustic characteristic.

| Acoustic property | $R^2$ | *p* |
|---|---|---|
| ratio.poss.act | .542 | <.001 |
| t.c.c | .700 | <.001 |
| all.j.c | .784 | <.001 |

Table 2
Multiple regression analysis modelling MDS coordinate values from acoustic characteristics: *y*-dimension. See Appendix A for a description of each acoustic characteristic.

| Acoustic property | $R^2$ | *p* |
|---|---|---|
| t.c.v | .256 | .012 |
| no.double | .507 | .001 |
| j.31.40 | .683 | <.001 |
| j.ratio.cv.cc | .777 | .011 |

Table 3
Multiple regression analysis modelling MDS coordinate values from acoustic characteristics: *z*-dimension (significant at $p = .06$). See Appendix A for a description of each acoustic characteristic.

| Acoustic property | $R^2$ | *p* |
|---|---|---|
| all.unit.err | .153 | .059 |
| j.less.20 | .341 | .012 |
| t.v.v | .490 | .003 |
| final.dip | .596 | .001 |
| t.v.c | .676 | .001 |
| incorr.str.int | .784 | <.001 |

analysis showed that the addition of each new predictor made a significant contribution to the model for each dimension ($p = .01$ or less for each addition); Durbin–Watson analysis confirmed independence of errors.

An examination of these correlations with respect to the general categories and distinctions identified in the auditory and visual analysis described above shows that all three of the dimensions of the MDS map were accounted for by some aspect of join quality/quantity, and by some aspect of unit appropriateness, while only one dimension was accounted for by any aspect of intonation/prosody. However, each dimension was accounted for by very different aspects of, and amounts of, each of these general acoustic characteristics, as described below.

### 3.2.1. The x-dimension: overall join quality/quantity

As can be seen in Table 1, the distribution of points along the *x*-dimension was predominantly accounted for by measures of join quality and/or quantity, namely the number of joins (normalised for the number of possible joins, as this will vary across utterances), and average join cost across an utterance. Together these two measures can be thought of as capturing some sort of high-level or global impression of the nature (i.e., number and overall quality) of the joins in an utterance. Additionally variability in the *x*-dimension was accounted for by one aspect of unit appropriateness: the average target cost of consonant–consonant diphones.

### 3.2.2. The y-dimension: join distribution and detectability

Table 2 shows that, like the *x*-dimension, variation in the *y*-dimension was accounted for predominantly by measures of join quality/quantity. However, it appears that where the *x*-dimension appeared to reflect some sort of overall impression of join behaviour, the *y*-dimension reflects more fine-grained detail regarding the distribution and detectability of joins in an utterance. First, variation in the *y*-dimension is predicted by the number of joins at 2-diphone intervals in an utterance (as reflected by "number of double units"). This is clearly some sort of measure of join number/distribution, and may also reflect join detectability. For example, this 2-diphone interval could, depending on the word, be roughly equivalent to a mora or syllable (for example, in the word "hippopotamus," a join every 2-diphones would place a join either in the middle of every consonant, or in the middle of every vowel). Given the perceptual importance of mora- and syllable-like units (see e.g., Cutler and Otake, 1994), the presence of joins (and thus potentially join discontinuities) at these intervals might be particularly salient to listeners. Second, variation along the *y*-dimension is accounted for by the proportion of joins with a join cost of between .31 and .40, which is the second highest (i.e., second worst quality) of the four divisions of join costs made in this study, and is thus simply a measure of the number of relatively poor quality joins in an utterance, normalised for the overall number of joins in the utterance. It is interesting to note that it is not the highest join costs (i.e., the join costs that reflect the least smooth, most audible joins) that are the most perceptually important in this dimension. This could, however, have something to do with the distribution of the join costs across the artificial levels utilised in this study. The first two divisions of join costs, covering the two lowest costs (less than .20 and .21–.30) appear in roughly equal proportion across all 24 of the utterances used in this study (40.58% and 41.79% respectively). The highest level of join cost (more than .41), on the other hand, is rarely seen (only 4.79% of all joins had this cost), Thus it is possible that joins in the range .31–.40, which accounted for 12.83% of all joins in this study, might be more perceptually noticeable. Variation along the *y*-dimension is also predicted by the segmental location of joins (as reflected by "ratio of joins in consonants to joins in vowels"). This last correlation is not surprising: studies have found that listeners' ability to detect discontinuities at join points depends on the segmental context of the join in question, with joins in vowels more detectable than joins in consonants (e.g. Klabbers and Veldhuis, 2001; Syrdal, 2001). Finally, again, like the *x*-dimension, there is also a role for appropriateness of units (specifically consonant–vowel units) in accounting for variability in the *y*-dimension.

#### 3.2.3. The z-dimension: unit appropriateness and prosody

Table 3 shows that in contrast to the *x*- and *y*-dimensions, variance along the *z*-dimension was accounted for more by measures of unit appropriateness than by measures of join quality/quantity. First, variance in the *z*-dimension is accounted for by a characteristic ("unit errors") that combines number of transcription/pronunciation errors, number of inappropriate diphones, number of Festival-identified "bad units", and number of missing diphones. This measure reflects those unit errors in which segmental information is outright incorrect or missing, rather than simply a poor match to the target. It should be pointed out that the salience of this type of segment error might not be strictly to do with segmental incorrectness: it is probable that all of these errors would additionally cause some sort of discontinuity in the signal at a join point, thus possibly making these errors particularly perceptible. Second, the *z*-dimension was also predicted by the average target cost across an utterance for vowel-initial units (vowel–vowel diphones and vowel–consonant diphones), potentially reflecting something to do with the distribution or detectability of unit errors.

An aspect of join cost does, however, play some role in accounting for variance along the *z*-dimension, specifically the proportion of joins with a join cost of less than .20, which is the lowest (i.e., best quality) of the four divisions of join costs made in this study. It is possible that this aspect of join cost in fact corresponds with the large number of unit-appropriateness predictors of the *z*-dimension, because joins appear to play a much larger overall role in perceived naturalness than units. It may thus be necessary to have low join costs (i.e., good, or less detectable joins) in order to be able to judge the appropriateness of segmental units.

Finally, the *z*-dimension was the only dimension that was accounted for by any aspect of intonation and/or prosodic appropriateness. First, the *z*-dimension was accounted for by a measure indicating whether or not the final diphone in a synthetic utterance was taken from utterance-final position in the source utterance (that is, whether or not the final diphone carried the appropriate utterance-final stress and intonation pattern). This is unsurprising given the importance of utterance-final prosody (which can include any of the following: lengthening of syllables, pitch declination, pausing) in the perception of phrase boundaries (Wightman et al., 1992; see Fisher and Tokura, 1996, for a review). Second, a small percentage of the variation in the *z*-dimension was also predicted by a measure of the number of instances of incorrect stress/intonation, across an utterance. These two measures capture both a global/high-level picture of the prosody errors across an utterance, and a more detailed picture of one important aspect of prosody, namely utterance final prosody.

### 3.3. Regression analysis: accounting for the acoustic characteristics

The linear regression described in Section 3.2 was performed with the *x*-, *y*- and *z*-axes as the response variables because these were the axes on which subjects distinguished the synthetic speech stimuli and, as noted above, the main aim of the current study was to devise an acoustically valid account of the listeners' perceptual evaluation behaviour. However, it is also possible to analyse the space from an alternate perspective, and perform linear regression with each acoustic characteristic in turn as a response variable and the *x*, *y* and *z* values as input variables. The results of such an analysis can be interpreted as follows: MDS analysis demonstrates that the listeners who participated in the current experiment made use of certain psychoacoustic characteristics of speech that correspond to the *x*-, *y*- and *z*-dimensions of the MDS space. Given that this is the case, if we were to present the same listeners with speech samples about which we knew nothing (with regard to acoustic characteristics), the *x*-, *y*- and *z*-dimensions of the current MDS space could be used to predict variation in certain of the acoustic characteristics of the new speech.

The seven out of 52 acoustic characteristics for which a significant amount of variance can be accounted with a linear model consisting of a constant and the three axis values are reported in Table 4. The *x*-dimension contributes to the variation in the largest number of acoustic characteristics. Five of these are related in some way to joins, namely the number of joins in an utterance (normalised for the number of possible joins), the length (in diphones) of the longest contiguous section of database speech in a synthetic utterance, the number of contiguous sections of database speech that were four diphones long or longer in a synthetic utterance, the number of joins at 2-diphone intervals, and the number of joins with costs of more than .40. These all indicate the strong relationship between the *x*-dimension and measures of join acceptability. In particular these measures correspond predominantly to number of joins and/or join location across utterances, with one measure representing severity or detectability of join (number of joins with costs of more than .40). Note, however, that (as discussed in Section 3.2.2) the number of joins at 2-diphone intervals could also be considered to be a measure of join detectability as well as join number/distribution. The *x*-dimension also accounts for variation in two measures of unit acceptability: average target cost across an utterance, and the average

Table 4
Multiple regression analysis modelling acoustic characteristics from MDS coordinate values. In addition to the significances shown, the intercept is significant for all models at $p < .001$.

| Response variable | $R^2$ | $p$ | Signif. level of axes | | |
|---|---|---|---|---|---|
| | | | *x* | *y* | *z* |
| ratio.poss.act | 0.5606 | 0.000 | 0.001 | – | – |
| longest | 0.4957 | 0.003 | 0.001 | – | – |
| no.long | 0.4244 | 0.011 | 0.01 | – | – |
| no.double | 0.4617 | 0.005 | 0.05 | 0.01 | – |
| j.more.41 | 0.4850 | 0.004 | 0.01 | 0.05 | – |
| all.t.c | 0.4295 | 0.009 | 0.01 | – | – |
| t.c.c | 0.4445 | 0.007 | 0.01 | – | – |

target cost across an utterance for consonant–consonant diphones.

The *y*-dimension contributes to the variation in only two of the above characteristics, both related to joins, and more specifically, both related in some way to join detectability: the number of joins at 2-diphone intervals, and the number of joins with costs of more than .40. The *z*-dimension did not make any contribution to the variation in the analysed acoustic characteristics at the significance levels reported here.

Interestingly, the spread of acoustic characteristics that can be predicted by the *x*-, *y*- and *z*-dimensions of the MDS space correspond fairly closely to those characteristics that were identified by means of the visual/auditory and cluster analysis as perceptually important. The groupings within the MDS space indicated that listeners were sensitive to both the number of joins in an utterance, and the quality of joins across an utterance, and additionally appeared sensitive to the number of poor quality units across an utterance. Similarly, the regression analysis performed in the current section has identified both number/distribution of joins and join severity/detectability as perceptually important, with measures of unit quality playing a slightly smaller perceptual role. Neither the visual/auditory and cluster analysis nor this regression analysis showed any influence of intonation/prosody.

## 4. Conclusions

The aim of the current study was to make use of multidimensional scaling (MDS) techniques to help identify those acoustic characteristics of unit-selection synthetic speech that might play a role in listeners' evaluation of such speech.

The first point to note is the overall importance of joins. For all three dimensions of the MDS space at least 20% of the variance can be accounted for by some aspect of joins; for some of the dimensions this percentage is much higher. It is perhaps unsurprising that joins should receive the highest perceptual weight, as in natural speech the presence of audible "pops" and "clicks" in the speech stream would be extremely unexpected. What is more surprising is the fact that both number of joins in an utterance and quality of joins make independent contributions to listeners' perception of an utterance's naturalness: this is evident from all three types of analysis carried out in this study (visual/auditory/cluster analysis; regression analysis to account for the MDS map; regression analysis to predict acoustic characteristics). In addition, listeners appear sensitive both to the general or overall behaviour of joins across an utterance (such as captured by "average join cost") as well as to more specific, context-dependent behaviour of joins (such as captured by "ratio of consonant-located joins to vowel-located joins"). Interestingly, this last finding replicates earlier work showing that listeners are sensitive to differences in joins in different phonetic contexts (Klabbers and Veldhuis, 2001; Syrdal, 2001), but expands the finding to full sentences (the original work was done on isolated syllables). It is clear from the current study, therefore, that it is not simply the case that listeners find unexpected spectral discontinuities in the speech stream extremely salient. Instead, it appears that listeners are highly sensitive to many aspects of joins, and differentiate between them at very subtle levels.

Measures of unit acceptability do not appear to play as all-encompassing a role as those of join appropriateness. The *x*- and *y*-dimensions of the MDS space were each accounted for by only one unit-related measure (as compared to two and three join-related measures, respectively), and only the *x*-dimension predicted any unit-related acoustic characteristics in the regression analysis to predict acoustic characteristics (and only two unit-related measures as compared to five join-related measures). Half of the acoustic characteristics that accounted for variation along the *z*-dimension were unit-related, but note that correlations with this dimension were calculated at a less stringent $p = .06$. However, those acoustic characteristics that were found to account for the *x*-, *y*-, and *z*-dimensions each accounted for a relatively large proportion of the variability along these axes. Thus, although we do not find a large number of different unit-related acoustic characteristics that correlate with the MDS space, it is clear that unit acceptability is highly perceptually important; this is also clear from the visual/auditory/cluster analysis. In particular, it appears that different dimensions of the MDS space seem to be affected differently by specific and context-dependent aspects of unit acceptability, namely target cost in each type of diphone (e.g., consonant–consonant, consonant–vowel, etc.), as well as more general measures such as average unit cost and number of auditorily detectable unit errors. One point that is interesting to note is the relationship between join costs and unit costs, particularly as seen in the results of the regression analysis for the *y*-axis. Variation along this axis is predicted by high quality joins and numerous measures of unit appropriateness. This grouping of acoustic cues suggests that in order to be able to judge unit appropriateness, there must be few distractions from inappropriate joins: that is, that listeners' overall heavier weighting of joins "washes out" much of the perceptual effect of unit appropriateness.

Finally, it is important to note the very small role of stress/intonation appropriateness in listeners' judgements of the naturalness of the 24 utterances used in this study: it accounts for a little more than 10% of the variance in the *z*-dimension (and again it should be noted that correlations with this dimension were calculated at $p = .06$), and no measures of intonation/prosody were predicted by the axes of the MDS space. It is possible that the way in which stress/intonation appropriateness was calculated here does not actually reflect listeners' perception of stress/intonation in synthetic speech. However, given the dominance of join cost measures and unit appropriateness measures in these correlations, it is equally possible that stress/intonation

appropriateness is simply a characteristic of synthetic speech that by default receives much less weight than other acoustic characteristics. Certainly this result concurs with research from previous studies of synthetic speech in which intonation variation was explicitly controlled (Hirst et al., 1998; Jilka et al., 2003; Syrdal and Jilka, 2004; Vainio et al., 2002). The results of the current study also support research on the perception of natural speech such as that by Bradlow et al. (1999). These authors found that listeners are more able to encode and remember variability in the speech signal relating to the acoustic characteristics of different talkers (which would include a great deal of acceptable segmental variability) than they are to encode variations in the amplitude in the speech (which would come under the heading of prosodic information). It is particularly interesting to note that the lighter weighting of intonation is not exclusive to synthetic speech. The reason for this almost certainly lies in the fact that a high degree of variability in intonation and prosody is acceptable in natural speech. For example, it has been demonstrated that if a speaker (either natural or synthetic) produces an utterance with intonation that does not quite meet the expected prototypical pattern for the context, this unexpected production may only result in a different semantic/pragmatic interpretation by the listener (rather than an interpretation of the speech as foreign-accented or unnatural, for example, Jilka, 2005). Unless the chosen intonation pattern is explicitly forbidden in a given context, listeners will only become aware of intonational deviations from prototypical patterns after repeated deviations.

In addition to investigating which acoustic characteristics of synthetic speech influence listeners' judgements of naturalness, the intention of this study was to gain a more complete understanding of the relationship between acoustic characteristics of synthetic speech and listeners' responses, in order to develop better subjective evaluation methodologies. Our findings can be summarised under three categories: join quality, stress/intonation, and segmental quality.

### 4.1. Join quality

It seems clear that listeners have little difficulty attending to discontinuities at unit joins—in fact, listeners appear to be highly sensitive to many different aspects of join quality. The high weight given by default to this particular aspect of synthetic speech means that subjective evaluations of join quality are straightforward: there is little interference from other aspects of the speech.

### 4.2. Stress and intonation

On the other hand, although listeners are aware of stress and intonation, these characteristics receive very little perceptual weight by default. Therefore, it would be inadvisable to ask listeners to evaluate possible improvements to aspects of the stress or intonation of a synthesiser without first training the listeners, or carefully designing the stimuli or the method of presentation in order to change listeners' default weighting patterns. Without such training or careful experimental design, it is highly likely that listeners' judgements of stress and intonation quality will be influenced by their much heavier weighting of (for example) the quality of joins produced by the system.

### 4.3. Segmental quality

Finally, it appears some caution should also be taken in the evaluation of unit or segmental quality. Although this aspect of synthetic speech is clearly more accessible to listeners than is intonation, it does not seem to receive as much perceptual weight as quality of joins. Additionally, listeners seem to perceive unit quality in terms of multiple different aspects of unit appropriateness, with particular emphasis on the type of unit (e.g., consonant–consonant, etc.). This should be considered carefully when designing methods to evaluate unit appropriateness.

### 4.4. Implications for the evaluation of synthetic speech

The findings summarised above should be taken into account when designing subjective methods of synthetic speech evaluation and, in particular, when asking listeners to evaluate specific sub- or supra-segmental aspects of such speech.

Objective methods for evaluating synthetic speech do not (yet) correlate very strongly with human judgements (Falk et al., 2008), although they can now be used to make broad judgements, such as separating out the systems in the Blizzard Challenge into three groups of 'good', 'average' and 'poor'. The problems with comparing synthetic speech to a natural reference were mentioned earlier; these problems are particularly acute in the case of supra-segmental properties. It also seems unlikely that single-ended objective methods (i.e., those which do not require a reference signal) will ever be able to successfully rate the supra-segmental quality of synthetic speech. Whilst it is to be hoped that objective measures continue to improve—and that our findings can help to achieve that—they do not currently offer a solution to the difficulties inherent in subjective evaluation of synthetic speech, and of supra-segmental aspects in particular. Reliable objective measures remain an attractive, though elusive, proposition (Möller and Falk, 2009). However, whilst they may not be able to completely replace human judgements, they do have a place within the system design and development cycle, offering the possibility of tests repeated many times on large amounts of material, which would be impractical for subjective testing. One specific example of this would be to use the knowledge of the weighting given to different acoustic cues provided by the current study for tuning the target

and join costs of a concatenative system. Another example would be to incorporate a model of speech perception or an objective speech quality measure into the objective function for training a statistical parametric system. Again, the information provided by the current study with regard to the weight given by human listeners to the acoustic characteristics of synthetic speech would be invaluable in the design of such a model or measure.

## Appendix A. List of acoustic analyses made on synthetic utterances and natural source utterances

I. Automatic analyses: general measures.

| Characteristic | Description |
|---|---|
| all.j.c | Average join cost across a synthetic utterance |
| all.t.c | Average target cost across a synthetic utterance |
| total.cost | Total cost (combination of overall join cost and overall target cost across a synthetic utterance) |
| msec | Duration of synthetic utterance, in ms |
| syllables | Duration of synthetic utterance, in number of syllables |

II. Automatic analyses: target measures.

| Characteristic | Description |
|---|---|
| t.v.v, t.v.c, etc. | Average target cost across a synthetic utterance for all targets in each of: vowel–vowel diphones; vowel–consonant diphones; consonant–consonant diphones; consonant–vowel diphones |
| t.less.10, t.11.20, etc. | A count, per synthetic utterance, of targets with a cost of: less than .10 (a good match between source and target); between .11 and .20 (a less good match between source and target); between .21 and .30 (a relatively poor match between source and target); more than .31 (a very poor match between source and target; note that these divisions are different from those given for join costs because the two costs differ in overall distribution) |
| bad.units | Number of "bad units" per synthetic utterance |
| miss.dip | Number of "missing diphones" |
| init.dip | Whether the initial diphone in the synthetic utterance was taken from utterance-initial position in the source |
| final.dip | Whether the final diphone in the synthetic utterance was taken from utterance-final position in the source |

III. Automatic analyses: join measures.

| Characteristic | Description |
|---|---|
| j.c.cons, j.c.vowel, etc. | Average join cost across a synthetic utterance for joins in each of: consonants; vowels; stops; fricatives; affricates; nasals; approximants; liquids; schwa-based vowels; diphthongs; all other vowels; silences |
| j.less.20, j.21.20, etc. | A count, per synthetic utterance, of joins with a cost of: less than .20 (good joins); between .21 and .30 (less good joins); between .31 and .40 (relatively poor joins); more than .40 (very poor joins; note that these divisions are different from those given for target costs because the two costs differ in overall distribution) |
| poss.joins | Possible number of joins (the total number of joins that would be present in a synthetic utterance if it was created completely from individual diphones, rather than from a combination of diphones and larger, multi-diphone units as would normally be the case) |
| act.joins | Actual number of joins |
| ratio.poss.act | Ratio of possible number of joins to actual number of joins |
| j.ratio.c.v.num | Ratio of number of joins in a consonant to number of joins in a vowel |
| j.ratio.c.v.c | Ratio of join cost for joins in a consonant to join cost for joins in a vowel |
| longest | Number of diphones in the longest unit (the longest unit was the unit with the highest number of sequential diphones from a single source utterance, in that utterance) |
| no.long | Number of "long" units (long units were classed as any unit from a source utterance with more than four sequential diphones) |
| no.single | Number of "single" units (a single unit was a source unit made up of a single diphone) |
| no.double | Number of "double" units (a double unit was a source unit made up of two sequential diphones) |
| single.row | Largest number of single units in a row |
| double.row | Largest number of double units in a row |
| all.unit.err | A count of the number of unit errors (sum of two automatically calculated measures: (i) number of bad units, (ii) number of missing diphones, and three manually calculated measures: (i) transcription/pronunciation errors, (ii) segmentation errors, (iii) inappropriate diphones) |

IV. Manual analyses.

| Characteristic | Description |
|---|---|
| trans.pron | A count of the number of transcription/pronunciation errors |
| segmentation | A count of the number of segmentation errors |
| incorr.unit | A count of the number of contextually inappropriate diphones |
| aud.joins | A count of the number of spectrally detectable joins |
| incorr.str.int | A count of the number of instances of incorrect stress or intonation |

## Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.specom.2010.10.003.

## References


Allen, P., Bond, C., 1997. Multidimensional scaling of complex sounds by school-aged children and adults. J. Acoust. Soc. Amer. 102 (4), 2255–2263.

Allen, P., Scollie, S., 2002. Stimulus set effects in the similarity ratings of unfamiliar complex sounds. J. Acoust. Soc. Amer. 112 (1), 211–218.

Bailly, G., Campbell, N., Mobius, B., 2003. ISCA special session: Hot topics in speech synthesis. <http://feast.his.atr.jp/synsig/euro-sig.pdf>.

Best, C.T., Morrongiello, B., Robson, R., 1981. Perceptual equivalence of acoustic cues in speech and non-speech perception. Percept. Psychophys. 29 (3), 191–211.

Black, A., Taylor, P.A., Caley, R., Clark, R., King, S., Richmond, K., 1997–2004. The Festival speech synthesis system. Available from CSTR <http://www.cstr.ed.ac.uk/projects/festival>.

Bradlow, A.R., Nygaard, L.C., Pisoni, D.B., 1999. Effects of talker, rate, and amplitude variation on recognition memory for spoken words. Percept. Psychophys. 61, 206–219.

Cerňak, M., Rusko, M., 2005. An evaluation of synthetic speech using the PESQ measure. In: Proc. Forum Acousticum, Budapest, pp. 2725–2728.

Cerňak, M., Rusko, M., Trnka, M., 2009. Diagnostic evaluation of synthetic speech using speech recognition. In: Proc. ICSV1, International Congress on Sound and Vibration, Kraków.

Chen, J., Campbell, N., 1999. Objective distance measures for assessing concatenative speech synthesis. In: Proc. Eurospeech'99, Sixth European Conference on Speech Communication and Technology, Budapest, Hungary, pp. 611–614.

Christensen, L.A., Humes, L.E., 1997. Identification of multidimensional stimuli containing speech cues and the effects of training. J. Acoust. Soc. Amer. 102 (4), 2297–2310.

Clark, R.A.J., 2003. Modelling pitch accents for concept-to-speech synthesis. In: Internat. Congress of Phonetic Sciences, Barcelona, Spain, pp. 1141–1144.

Clark, R.A.J., Dusterhoff, K.E., 1999. Objective methods for evaluating synthetic intonation. In: Proc. Eurospeech'99, Sixth European Conf. on Speech Communication and Technology, Budapest, Hungary, pp. 1623–1626.

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

Cutler, A., Otake, T., 1994. Mora or phoneme? Further evidence for language-specific listening. J. Memory Lang. 33, 824–844.

Expert Advisory Group on Language Engineering Standards, 1996. Evaluation of natural language processing systems: Final report. <http://www.issco.unige.ch/projects/ewg96/ewg96.html>.

Falk, T.H., Moeller, S., Karaiskos, V., King, S., 2008. Improving instrumental quality prediction performance for the Blizzard Challenge. In: Proc. Blizzard Workshop, Brisbane, Australia.

Fisher, C., Tokura, H., 1996. Prosody in speech to infants: Direct and indirect acoustic cues to syntactic structure. In: Morgan, J.L., Demuth, K. (Eds.), Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition. Lawrence Erlbaum, Mahwah, NJ, pp. 343–363.

Francis, A.L., Kaganovich, N., Driscoll-Huber, C., 2008. Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. J. Acoust. Soc. Amer. 124, 1234–1251.

Garofolo, J.S., 1988. Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database. National Institute of Standards and Technology (NIST), Gaithersburgh, MD.

Gordon, P.C., Eberhardt, J.L., Rueckl, J.G., 1993. Attentional modulation of the phonetic significance of acoustic cues. Cogn. Psychol. 25, 1–42.

Hall, J.L., 2001. Application of multidimensional scaling to subjective evaluation of coded speech. J. Acoust. Soc. Amer. 110, 2167–2182.

Hazan, V., Barrett, S., 2000. The development of phonemic categorisation in children aged 6–12. J. Phonetics 28, 377–396.

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

Hazan, V., Simpson, A., Huckvale, M., 1998. Enhancement techniques to improve the intelligibility of consonants in noise: Speaker and listener effects. In: ICSLP, Sydney, Australia, pp. 2163–2167.

Hirst, D., Rilliard, A., Aubergé, V., 1998. Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis. In: Proc. ESCA/COCOSDA Workshop on Speech Synthesis'98, Jenolan Caves, Blue Mountians, Australia.

ITU-T Recommendation P.85, 1994. A method for subjective performance assessment of the quality of speech output devices. International Telecommunications Union publication.

Iverson, P., Kuhl, P.K., Akahane-Yamada, R., Tohkura, D.E.Y., Kettermann, A., Siebert, C., 2002. A perceptual interference account of acquisition difficulties for non-native phonemes. Cognition 87, B47–B57.

Iverson, P., Hazan, V., Bannister, K., 2005. Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/–/l/ to Japanese adults. J. Acoust. Soc. Amer. 118, 3267–3278.

Jilka, M., 2005. Exploration of different types of intonational deviations in foreign-accented and synthesized speech. In: Proc. Interspeech-2005, Lisbon, Portugal, pp. 2393–2396.

Jilka, M., Syrdal, A., Conkie, A., Kapilow, D., 2003. Effects on TTS quality of methods of realizing natural prosodic variations. In: Proc. ICPhS, International Congress of Phonetic Sciences, Bercelona, Spain, pp. 2549–2552.

Jusczyk, P.W., 1997. The Discovery of Spoken Language. MIT Press, Cambridge, Massachusetts.

Klabbers, E., Veldhuis, R., 1998. On the reduction of concatenation artefacts in diphone synthesis. In: Proc. ICSLP'98, 5th Internat. Conf. on Spoken Language Processing, Sydney, Australia, pp. 1983–1986.

Klabbers, E., Veldhuis, R., 2001. Reducing audible spectral discontinuities. IEEE Trans. Speech Audio Process. 9.

Kreiman, J., Gerratt, B.R., 1998. Validity of rating scale measures of voice quality. J. Acoust. Soc. Amer. 104, 1598–1608.

Kreiman, J., Gerratt, B.R., 2000. Sources of listener disagreement in voice quality assessment. J. Acoust. Soc. Amer. 108, 1867–1876.

Kreiman, J., Gerratt, B.R., 2004. Perceptual relevance of source spectral slope measures. J. Acoust. Soc. Amer. 115, 2609.

Kreiman, J., Gerratt, B.R., Ito, M., 2007. When and why listeners disagree in voice quality assessment. J. Acoust. Soc. Amer. 122 (4), 2354–2364.

Kruskal, J.B., Wish, M., 1978. Multidimensional Scaling. Sage University Paper series on Quantitative Applications in the Social Sciences. Sage Pubns., Beverly Hills and London.

Lamel, L.F., Kassel, R.H., Seneff, S., 1989. Speech database development: Design and analysis of the acoustic–phonetic corpus. In: Proc. Speech I/O Assessment and Speech Databases, Noordwijkerhout, The Netherlands, pp. 2161–2170.

Marozeau, J., de Cheveigné, A., McAdams, S., Winsberg, S., 2003. The dependency of timbre on fundamental frequency. J. Acoust. Soc. Amer. 114, 2946–2957.

Mayo, C., Turk, A., 2004. Adult-child differences in acoustic cue weighting are influenced by segmental context: Children are not always perceptually biased toward transitions. J. Acoust. Soc. Amer. 115, 3184–3194.

Mayo, C., Turk, A., 2005. The influence of spectral distinctiveness on acoustic cue weighting in children's and adults' speech perception. J. Acoust. Soc. Amer. 118, 1730–1741.

Mayo, C., Clark, R.A.J., King, S., 2005. Multidimensional scaling of listener responses to synthetic speech. In: Proc. Interspeech 2005, Lisbon, Portugal.

Möller, S., Falk, T.H., 2009. Quality prediction for synthesized speech: Comparison of approaches. In: Proc. NAG/DAGA 2009, Rotterdam, pp. 1168–1171.

Nittrouer, S., 2004. The role of temporal and dynamic signal components in the perception of syllable-final stop voicing. J. Acoust. Soc. Amer. 115, 1777–1790.

Plumpe, M., Meredith, S., 1998. Which is more important in a concatenative text to speech system—pitch, duration, or spectral discontinuity? In: Proc. ESCA/COCOSDA Workshop on Speech Synthesis'98, Jenolan Caves, Blue Mountians, Australia.

Rabinov, C.R., Kreiman, J., Gerratt, B.R., Bielamowicz, S., 1995. Comparing reliability of perceptual ratings of roughness and acoustic measures of jitter. J. Speech Hear. Res. 38, 26–32.

Schnieder, W., Eschman, A., Zuccolotto, A., 2002. E-Prime User's Guide; E-Prime Reference Guide. Psychology Software Tools, Inc., Pittsburgh, PA.

Stylianou, Y., Syrdal, A.K., 2001. Perceptual and objective detection of discontinuities in concatenative speech synthesis. In: Proc. ICASSP, Internat. Conf. on Acoustics, Speech and Signal Processing, Salt Lake City, Utah.

Syrdal, A., 2001. Phonetic effects on listener detection of vowel concatenation. In: Proc. Eurospeech 2001, 7th European Conference on Speech Communication and Technology, Aalborg, Denmark, pp. 979–982.

Syrdal, A., Jilka, M., 2004. Acceptability of variations in question intonation in natural and synthesised American English. J. Acoust. Soc. Amer. 115 (5), 2543(A).

Turk, A., Nakai, S., Sugahara, M., 2006. Acoustic segment durations in prosodic research: A practical guide. In: Sudhoff, S., Lenertova, D., Meyer, R., Pappert, S., Augurzky, P., Mleinek, I., Richter, N., Schliesser, J. (Eds.), Methods in Empirical Prosody Research. De Gruyter, Berlin, pp. 1–28.

Vainio, M., Järvikivi, J., Werner, S., 2002. Effect of prosodic naturalness on segmental acceptability in synthetic speech. In: Proceedings IEEE 2002 Workshop on Speech Synthesis, Santa Monica, California.

Vepa, J., King, S., 2004. Join cost for unit selection speech synthesis. In: Alwan, A., Narayanan, S. (Eds.), Text to Speech Synthesis: New Paradigms and Advances. Prentice-Hall, Upper Saddle River, NJ.

Wardrip-Fruin, C., 1982. On the status of temporal cues to phonetic categories: Preceding vowel duration as a cue to voicing in final stop consonants. J. Acoust. Soc. Amer. 71, 187–195.

Wardrip-Fruin, C., 1985. The effect of signal degradation on the status of cues to voicing in utterance-final stop consonants. J. Acoust. Soc. Amer. 77 (5), 1907–1912.

Watson, J., 1997. Sibilant-vowel coarticulation in the perception of speech by children with phonological disorder. Ph.D. Dissertation, Queen Margaret College, Edinburgh.

Wightman, C., Shattuck-Huffnagel, S., Ostendorf, M., Price, P., 1992. Segmental durations in the vicinity of prosodic phrase boundaries. J. Acoust. Soc. Amer. 91, 1707–1717.

Wouters, J., Macon, M.W., 1998. A perceptual evaluation of distance measures for concatenative speech synthesis. In: Proc. ICSLP'98, 5th Internat. Conf. on Spoken Language Processing, Sydney, Australia, pp. 2747–2750.

Zen, H., Tokuda, K., Kitamura, T., 2007. Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences. Comput. Speech Lang. 21 (1), 153–173.