Multidimensional scaling of listener responses to synthetic speech

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

Centre for Speech Technology Research
University of Edinburgh, UK

catherin@ling.ed.ac.uk, robert@cstr.ed.ac.uk, simon.king@ed.ac.uk

Abstract

The move to unit-selection in speech synthesis has resulted in system improvements 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.

This paper describes a pilot study which aims to evaluate multidimensional scaling (MDS) as a possible method of determining what acoustic characteristics of synthetic speech influence listeners' judgements of the naturalness of the speech. Using distance measures (either real or perceived distances), MDS techniques represent stimuli as points in n-dimensional space. The space is configured so that similar stimuli are close together, while different stimuli are farther apart. Additionally, the dimensions of the space correspond to characteristics of the stimuli which influenced the perceived distances.

Our results indicate that MDS techniques should be a useful tool in understanding the complex psychoacoustic processes that listeners undergo when evaluating synthetic speech. This method has allowed us to identify a number of cues that appear to be particularly perceptually salient to listeners evaluating synthetic speech naturalness, namely *prosodic cues* (in terms of duration and/or intonation) and *segmental or unit level cues* (in terms of appropriateness of units, or number of units).

1. Introduction

Great progress has been made recently in speech synthesis, most notably the move to unit-selection, and system improvements are now being made at subtle sub- and supra-segmental levels (e.g., unit joins, intonation). Human perceptual evaluation of such subtle improvements often requires listeners to attend to just one dimension of a complete synthesis system. However, numerous studies have found that raters are often adversely affected by dimensions of the signal other than those they have been asked to rate. For example, listeners' judgements of intonation naturalness have been shown to be influenced by segmental quality [1, 2], while intonation appropriateness has been found to impact on perceived segmental quality [2].

A number of studies, both from the field of auditory evaluation, and from the wider field of auditory perception, point to possible reasons for these findings. First, it appears that when faced with complex acoustic stimuli that vary along multiple dimensions, listeners find it difficult to focus on just one dimension. For example, it has been found that listeners are much less able to rate intonation consistently when it varies simultaneously with many other acoustic dimensions than when intonation is the only aspect of the stimulus set to be varied [3]. Furthermore, it is not simply the case that listeners give equal attention or perceptual "weight" to all available acoustic information. Instead listeners give more weight to some dimensions than others. For example, the addition of synthetic intonation to natural speech segments was found to be more detrimental to listeners' quality ratings than was the addition of synthetic segment duration [4], suggesting that for this listening situation, intonation was weighted more heavily than segment duration.

Evidence from across speech perception and general auditory perception [5, 6, 7, 8, 9] indicates that listeners' hierarchies of weighting can differ depending on the segmental and acoustic context of the stimuli (e.g., speech versus non-speech, natural speech versus synthetic speech, first language versus second language, etc.). Unfortunately, no one has examined the acoustic dimension weighting behaviour of listeners when rating synthetic speech. It is therefore unclear whether listeners are, for example, consistently more influenced in a speech synthesis rating task by segmental quality, or by appropriateness of intonation. The goal of the current line of research, therefore, is to determine the pattern of weights listeners give to available acoustic dimensions (both sub- and supra-segmental) when rating synthetic speech.

This paper describes a pilot study which examines the suitability of multidimensional scaling (MDS) [10] for identifying the main acoustic dimensions to which listeners attend when rating synthetic speech. MDS techniques use measures of proximities (either real distances or perceived psychophysical distances) between objects to derive a *stimulus space*, in which the distances between stimuli in the space correspond to the proximity values (similar stimuli are placed close together; dissimilar stimuli are placed further apart). Additionally, the dimensions that make up the stimulus space correspond to the dimensions used most heavily by the listeners to make their proximity judgements. Subsequent analysis of these dimensions can reveal the physical or psychophysical characteristics of the stimuli on which proximity judgements are made. MDS techniques have been used successfully to determine the underlying characteristics responsible for perceptual decisions in numerous auditory domains: e.g., complex non-speech sounds [11, 5], coded speech [12], segmental contrasts [13], voice quality [14], and musical timbre [15]. In the case of perceptual evaluation of synthetic speech, the use of MDS techniques and subsequent analysis of the resulting stimulus space should allow for the identification of those acoustic cues which most influence listeners' perception of "naturalness" in such speech.

Table 1: *Timit sentences used to create synthetic utterances*

Utt. No.	Sentence	Duration (sec)
1	As a precaution, the outlaws bought gunpowder for their stronghold.	3.8
2	Her auburn hair reminded him of autumn leaves.	2.6
3	They remained lifelong friends and companions.	3.0
4	Curiosity and mediocrity seldom coexist.	2.8
5	The easygoing zoologist relaxed throughout the voyage.	3.3
6	Biologists use radioactive isotopes to study microorganisms.	4.1
7	Employee layoffs coincided with the company's reorganization.	3.5
8	Who took the kayak down the bayou?	1.9

2. Method

2.1. Stimuli

To obtain a set of utterances which covered a range of qualities, 8 consecutive sentences from the text of the TIMIT database [16] were chosen at random to create 8 synthetic utterances (see Table 1 for complete list of sentences). The sentences ranged from 9 syllables to 22 syllables in length; the resulting synthetic utterances ranged from 1.9 sec to 4.1 sec.

The utterances were synthesised using the Festival 1.96 multisyn engine with a female, RP English voice (cstr_rpx_nina_multisyn). The utterances were not manipulated during or after synthesis to create more or less natural sounding utterances. However, analysis of the perceptual results, and informal post-test questioning of participants, indicates that they covered a range of perceived naturalness.

2.2. Participants

Eight adults ranging in age from 27 years to 35 years (average age: 33 years) took part in this perceptual evaluation experiment. All participants were native speakers of English and all reported themselves as being free from speech/language disorders. All participants were at least somewhat experienced with listening to synthetic speech.

2.3. Procedure

All participants were tested individually in a quiet room. The stimuli were presented over closed-back headphones (Sennheiser PX 200, frequency response 10-21000 Hz), with volume set constant at a comfortable listening level. Testing took place in one 40 min session, with two short breaks part way through testing.

The stimuli were presented in pairs. Presentation of these pairs of utterances was controlled by a suite of computer software [17]. The listener's task was to indicate, by typing in a response, whether the two utterances in a pair were 'similar' or 'different' in terms of their *naturalness*. The participants were instructed that they should ignore the lexical content of the utterances, and instead concentrate on the naturalness, that is, how much like 'real speech' the utterances sounded. Importantly, the participants were *not* instructed to listen to any one acoustic characteristic of the stimuli, or to any specific psychoacoustic construct (e.g., 'listening effort', 'pleasantness', 'pronunciation' etc) such as have been used in previous evaluation studies e.g., [18]. The task was simply to make a simple binary decision about the degree of similarity in naturalness of each pair of stimuli; MDS analysis should then derive the underlying physical or psychoacoustic characteristics on which these binary decisions were made.

Before testing, the participants were given an opportunity to listen to examples of the type of synthetic speech to be used in the test. Three pairs of utterances illustrating extreme examples of 'similar' and 'different' pairs were played to the listeners. These stimuli were synthesised using the same female RP voice as that used to synthesise the pre-test and main test utterances, however the utterances designed to illustrate extreme examples of *unnatural* synthetic speech were synthesised from a database of only 400 rather than 2000 sentences, to deliberately degrade the resulting stimuli.

Following this familiarisation period, a pre-test was administered to ensure that all participants understood the task. This pre-test consisted of synthetic versions of 9 Timit sentences, presented in pairs. These utterances were synthesised using the same female, RP voice and using the full 2000 sentence database as the main test utterances. None of the sentences were the same as those presented in the main test, in terms of lexical content. Nine of the possible pairs of utterances were presented, in random order.

The main test consisted of each of the 8 utterances paired with every other utterance, presented 6 times each (3 times in each order, i.e. AB, and BA), resulting in 168 pairs of utterances. The interval between the presentation of each member of the pair was 5000 msec (onset to onset). Responses were not timed and the presentation of the next pair of utterances began 2000 msec following the entry of a response. The 168 pairs of utterances were randomised for presentation. Breaks were given following the presentation of the 56th and the 112th pairs; the duration of these breaks was controlled by the participants.

3. Analysis

Listeners' 'similar' and 'different' responses were compiled into a dissimilarity matrix in which each cell in the matrix represented the number of times an utterance pair had received the label 'different.' Multidimensional scaling of the similarity matrix was carried out by means of SPSS (Version 11.5.0). Subsequent visual and auditory analysis of the configuration of the resulting stimulus space was carried out by the three authors. This visual and auditory analysis was confirmed by means of cluster analysis techniques.

4. Results

Multidimensional scaling indicates that it is appropriate to represent the results in three dimensions, as illustrated in Figure 1. With this number of dimensions a relatively high proportion of the variance in the data is accounted for (RSQ=0.968) with a fairly low level of residual stress (0.05004).

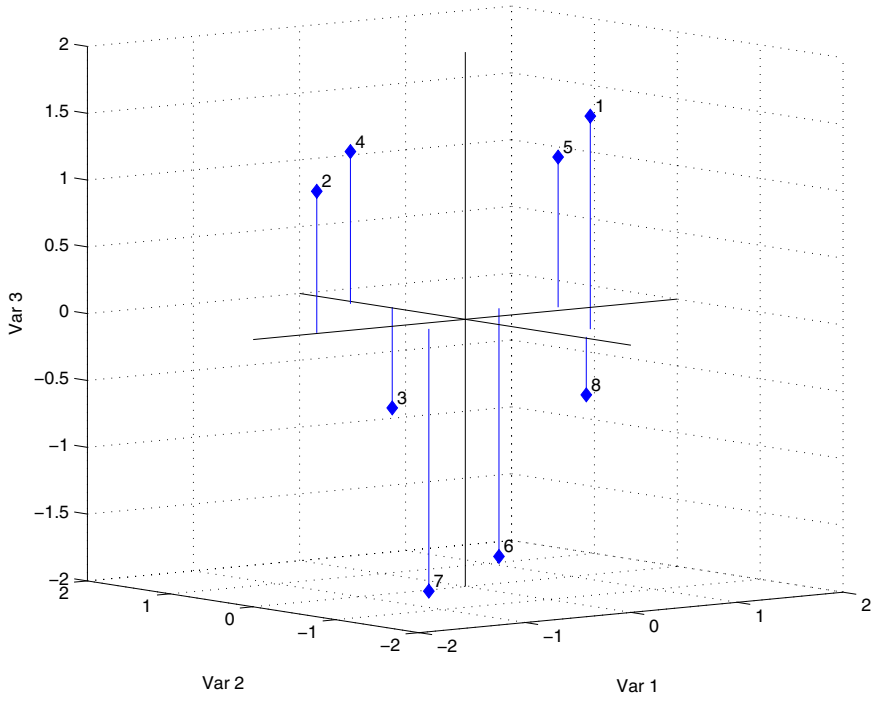

Figure 1: *Three dimensional MDS map from dissimilarity judgements of 8 synthetic speech utterances by 8 listeners. Numbers correspond to the number of the utterance as listed in Table 1 and as discussed in the text.*

Combined auditory and visual analysis of the configuration of the data indicates, first, that listeners perceived the utterances on a graded scale, with two fairly natural sounding utterances (Utterance 7 and Utterance 6) on one end of this scale, and a range of utterances (Utterances 2, 4, 5 and 1) at the other end of the scale.

Further analysis shows that the data fall into three main clusters. An examination of these clusters allows for the identification of two main acoustic characteristics that seem to underlie listeners' similarity judgements. The first cluster, consisting of Utterance 7 and Utterance 6, includes the most natural sounding utterances in the stimulus set. The second cluster consists of Utterance 5 and Utterance 1, which both have fairly extreme errors in prosody (either duration, intonation, or both). Cluster three consists of Utterance 2 and Utterance 4, both of which predominantly contain errors which can be classified as occurring at the segmental or unit level: either inappropriate units have been chosen, resulting in poor joins at unit edges, or possibly too many units have been used to create the utterance. Utterance 8 is fairly natural sounding, with one major error of timing/prosody; its placement between Cluster 1 and Cluster 2 is therefore unsurprising. Similarly Utterance 3, which falls between Cluster 1 and Cluster 3, is quite natural, but appears to contain a few unit-level errors. Cluster analysis confirms these visual and auditory analyses, producing the same main clusters as indicated above. Utterance 3 was clustered with the two most natural utterances, Utterance 7 and Utterance 6; Utterance 8 formed a single cluster on its own. Readers are encouraged to listen to the stimuli in conjunction with their examination of Figure 1: audio files can be found at http://www.ling.ed.ac.uk/∼catherin/synthetic-speech.

5. Discussion

This study suggests that the use of multidimensional scaling techniques should indeed help to provide a better understanding of how listeners perceive synthetic speech. Our study asked listeners to make a simple binary decision regarding the degree of similarity between a pair of stimuli. Furthermore, the listeners were asked to judge degree of similarity only on one general dimension, i.e., 'naturalness.' Informal post-test questioning of the participants showed that this was perceived to be a very easy task (in comparison to tasks which require the use of rating scales, for example). However, despite this perceived ease, MDS techniques show that the participants were in fact performing a fairly complex task, making perceptual decisions on the basis of at least two (probably interacting) dimensions.

Our results show that MDS techniques provide a useful tool for identifying the 'hidden' physical or psychophysical dimensions on which perceptual decisions regarding synthetic speech are made. The visual, auditory and cluster analyses of the configuration of the utterances provided by MDS allowed us to hypothesise that listeners judge the naturalness of synthetic speech stimuli based on at least two main acoustic cues: the appropriateness of prosody, and the appropriateness, or number, of units selected for synthesis. Further MDS studies, in which different aspects of these two characteristics are deliberately manipulated, should allow for the identification of the more fine-grained acoustic cues that may be involved in perceived naturalness.

A better understanding of how listeners perceive synthetic speech should allow for the development and use of more appropriate auditory evaluation procedures. As noted above, perceptual evaluation of sub- and supra-segmental characteristics of synthetic speech can be hampered by the fact that listen-

ers are often influenced by alternate aspects of the speech signal. However, it has been shown that with certain presentation methods, it is possible to change listeners' default attention patterns. Perceptual weights given to various acoustic cues to both speech and non-speech stimuli have been shown to be changed by training [7], by manipulation of the stimuli to mask certain cues (e.g., simultaneous white noise, [19, 20] or reverberation [21]) or to enhance certain cues [22, 23], by manipulation of the distribution of the acoustic dimensions across the whole stimulus set [5], or by manipulation of listeners' conscious focus of attention (e.g., by presenting the rating task simultaneously with another task, [24]). It appears possible, therefore, that if listeners do not by default give adequate attention to the dimension under investigation, appropriate methods could be designed to cause listeners to re-focus their attention. Knowledge of which acoustic characteristics listeners *do* find most salient when rating the naturalness of such speech, should allow for the most appropriate of these presentation methods to be used.

The development and use of more appropriate auditory evaluation procedures should lead to more consistent and reliable subjective measures of synthetic speech quality. Developers of speech synthesis systems will thus be able to determine more accurately the perceived quality of their systems, and to make substantiated claims about this quality. Developers of speech-enabled applications, and users of such applications, will, in turn, be able to independently verify the quality of a synthesis system, and make better-informed decisions as to choice of system.

6. Acknowledgements

This study was funded by EPSRC grant EP/C53042X/1. Simon King is supported by EPSRC Advanced Research Fellowship GR/T04649/01.

7. References

[1] D. Hirst, A. Rilliard, and V. Aubergé, "Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis," in *ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.

[2] M. Vainio, J. Järvikivi, and S. Werner, "Effect of prosodic naturalness on segmental acceptability in synthetic speech," in *IEEE Workshop on Speech Synthesis*, Santa Monica, California, 2002.

[3] J. Kreiman and B. R. Gerratt, "Sources of listener disagreement in voice quality assessment," *JASA*, vol. 108, pp. 1867–1876, 2000.

[4] M. Plumpe and S. Meredith, "Which is more important in a concatenative text to speech system—pitch, duration, or spectral discontinuity?," in *ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.

[5] P. Allen and S. Scollie, "Stimulus set effects in the similarity ratings of unfamiliar complex sounds," *JASA*, vol. 112, no. 1, pp. 211–218, 2002.

[6] C. T. Best, B. Morrongiello, and R. Robson, "Perceptual equivalence of acoustic cues in speech and non-speech perception," *Percep. & Psychophys.*, vol. 29, no. 3, pp. 191–211, 1981.

[7] L. A. Christensen and L. E. Humes, "Identification of multidimensional stimuli containing speech cues and the effects of training," *JASA*, vol. 102, no. 4, pp. 2297–2310, 1997.

[8] C. Mayo and A. Turk, "Adult-child differences in acoustic cue weighting are influenced by segmental context: Children are not always perceptually biased toward transitions," *JASA*, vol. 115, pp. 3184–3194, 2004.

[9] S. Nittrouer, "The role of temporal and dynamic signal components in the perception of syllable-final stop voicing," *JASA*, vol. 115, pp. 1777–1790, 2004.

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

[11] P. Allen and C.-A. Bond, "Multidimensional scaling of complex sounds by school-aged children and adults," *JASA*, vol. 102, no. 4, pp. 2255–2263, 1997.

[12] J. L. Hall, "Application of multidimensional scaling to subjective evaluation of coded speech," *JASA*, vol. 110, pp. 2167–2182, 2001.

[13] P. Iverson, P. K. Kuhl, R. Akahane-Yamada, Diesch E., Y. Tohkura, A. Kettermann, and C. Siebert, "A perceptual interference account of acquisition difficulties for non-native phonemes," *Cog.*, vol. 87, pp. B47–B57, 2002.

[14] J. Kreiman and B. R. Gerratt, "Perceptual relevance of source spectral slope measures," *JASA*, vol. 115, pp. 2609, 2004.

[15] J. Marozeau, A. de Cheveigné, S. McAdams, and S. Winsberg, "The dependency of timbre on fundamental frequency," *JASA*, vol. 114, pp. 2946–2957, 2003.

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

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

[18] A. Sluijter, E. Bosgoed, J. Kerkhoff, E. Meier, T. Rietveld, A. Sanderman, and J. Terken, "Evaluation of speech synthesis systems for Dutch in tellecommunications systems," in *ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998.

[19] C. Wardrip-Fruin, "On the status of temporal cues to phonetic categories: Preceding vowel duration as a cue to voicing in final stop consonants," *JASA*, vol. 71, pp. 187–195, 1982.

[20] C. Wardrip-Fruin, "The effect of signal degredation on the status of cues to voicing in utterance-final stop consonants," *JASA*, vol. 77, no. 5, pp. 1907–1912, 1985.

[21] J. Watson, *Sibilant-Vowel Coarticulation In The Perception Of Speech By Children With Phonological Disorder*, Ph.D. thesis, Queen Margaret College, Edinburgh, 1997.

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

[23] V. Hazan and A. Simpson, "The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise," *Sp. Comm.*, vol. 24, pp. 211–226, 1998.

[24] P. C. Gordon, J. L. Eberhardt, and J. G. Rueckl, "Attentional modulation of the phonetic significance of acoustic cues," *Cog. Psy.*, vol. 25, pp. 1–42, 1993.