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