THE EFFECTS OF VOICE TYPE AND QUALITY ON THE INTELLIGIBILITY OF A TEXT-TO-SPEECH SYSTEM

J Brian Pickering
Speech & Knowledge Based Systems Research Group
IBM UK Scientific Centre, Athelstan House, St Clement Street,
Winchester, SO23 9DR

ABSTRACT

In developing a text-to-speech system, it is important to consider not only how intelligible the final output may be, but also how acceptable the synthetic voice is to the users of such a system. We might therefore go so far as to consider offering a range of different types of voice, some male, some female, to allow the user the flexibility they may require. It is not clear, however, just how different voices and different voice qualities may affect the intelligibility of the text-to-speech system. Nor, indeed, is it clear whether generating different voice types might involve simple modification of the basic synthetic voice offered, or require the extraction of a whole new parameter set. This paper addresses three basic issues: first, can we generate a number of different voices by modifying a single voice; second, how acceptable do native listeners judge those voices to be; and finally, what, if any, is the effect of changing the voice type on the intelligibility of the synthetic output.

INTRODUCTION

This paper describes part of the continuing assessment of our text-to-speech system, EXPRESS. In developing such a system, it is important to consider not only the overall intelligibility of the synthetic output, but also the acceptability of the synthetic voice itself. Indeed, it would be desirable to offer a range of different voice types and qualities, each of which would lead to the same general intelligibility. There are, however, a number of issues here. Can we automatically generate such a range of voice types without substantial reanalysis of real speech? Further, can we be sure that different voice types will not reduce the intelligibility of the system? In informal tests, we have found that listeners with little experience of synthetic speech might like a particular type of voice and regard the overall quality to be highly natural without being able to understand everything that is said. With this in mind, we decided to investigate (i) how we might produce a range of different synthetic voices; (ii) how acceptable this voices are, and (iii) how they affect the general intelligibility of the system.

In trying to assess intelligibility, we are faced with the problem of the type of test to choose. Previously [1,2,3], researchers have used standard psychological tests such as the modified and diagnostic rhyme tests. These certainly provide a rapid overall assessment of phoneme recognition. However, it is important not to overestimate the results in the general assessment of the synthetic speech. Essentially, these tests provide a handle on the shortcomings of the synthetic speech: if a given phoneme or group of phonemes is consistently misidentified, then the investigator knows he has a problem. On the other hand, they do not really indicate in any but the broadest terms how well the synthesis is doing: subjects are given a closed set of responses, which they know to contain the 'correct' answer, and therefore are not required to recognise the synthetic items they hear but rather to say what lexical item the synthetic word sounds most like. Notwithstanding such shortcomings, a modified rhyme test (MRT) seemed appropriate here precisely because it provides diagnostic information about where the system is going wrong. Such information is invaluable in attempts to improve the sound quality in general.

BACKGROUND

Pisoni and his fellow workers have used the MRT and other tests in an attempt to show how well MITalk performed [1]. These tests, however, were run with a single voice type. Pratt has since investigated a number of voice types and synthesizers, using a number of paradigms in order to quantify and compare the performance of a given system [2,3]. Further, he asked listeners to rate the naturalness, intelligibility and so forth of the test voices [2], and was able to show strong correlations between these subjective features. But this information was gathered in conjunction with a diagnostic rhyme test; subjects were therefore exposed to little more 'natural' speech than individual test items. Furthermore, the goal there was to compare and contrast different synthesis systems, rather than establish any link between voice type and performance. The present study sets out with this in mind.

We should consider, however, whether it is viable to attempt to synthesise a number of different voice types from a base set of synthesis parameters, or whether indeed it may be necessary to generate a series of parameter sets, one for each voice type. In an earlier study [4&5], we showed how an acceptable female voice could be generated automatically from a set of (copy) synthesis parameters for a male voice after appropriate normalising shifts for formant centre frequency and fundamental frequency. That study further showed how the acceptability of the synthetic voices and the perceived speaker category could be affected by the choice of different glottal pulse shapes to drive the synthesis. Based on this work, we might hope to be able to offer a range of different synthetic voices, and by careful manipulation of global parameters, generate a number of different voice types based on the original set.

EXPERIMENTAL PROCEDURE

Three different voices were generated: V1 is the default male voice of our text-to-speech system, and is the result of extracting appropriate acoustical parameters from a single male RP speaker; V2 is a female voice based on V1 after an appropriate normalisation; V3 is a second male voice, produced by applying the inverse normalisation shift to that used to generate V2, again to the default male voice V1.

V1 default male voice
V2 female voice derived from V1
V3 male voice derived from V1

The Synthetic Voices Offered

Each of these voices was generated with three basic global parameter settings: T1 is the optimum setting from [4]: the most natural glottal pulse shape from Rosenberg [5], plus a lip-radiation filter; T2 uses the lip-radiation filter with the
standard impulse-train type excitation offered by the synthesizer, the lip-radiation filter essentially preemphasises the synthetic output: T3 is the default synthesizer setting with no lip-radiation filter and the impulse-train excitation. Pairing each voice and each type, then, yields nine voice types. These were used to generate synthetic utterances and monosyllables on our own formant synthesizer [6].

Two tests were run: the first consisted of three utterances synthesised entirely by EXPRESS with the nine voice types. The utterances were: This is the BBC World Service. She sells sea-shells on the seashore, and Hello! How are you? The last utterance was used for comparison with the earlier study with copy synthesis parameters [4] and had the original intonation contour superimposed; for the others, a synthetic contour was generated by the text-to-speech system. Taking all three utterances together, a fair range of speech sounds, including stops and fricatives, was covered. The second test was of the MRT type: the eighteen consonants /p b t d k g f v s l z f r w m n/ occurred in syllable-initial positions; while the sixteen consonants /p b t d k g f v s l z f r w m n/ appeared in syllable-final positions. In both cases, the syllable nucleus was the vowel /i/; this yields many minimal sets which can be used in an MRT type test.

All subjects had little experience with synthetic speech. In the first test, listeners heard the 27 synthetic utterances (3 voices × 3 types × 3 utterances) in random order ten times and with a 4.5-second interval between each test item. They were asked to decide whether the voice sounded most like a man, woman or child, and to judge how acceptable the synthetic voice sounded, from 1 (unacceptable) to 6 (highly acceptable). The subjects were told that they would be listening to synthetic speech. In the second test, subjects heard the 306 monosyllables (18 × 16 consonants × 3 voices × 3 types) in random order, with a 3.50-second interval between successive items: the syllable-initial and syllable-final items were tested separately. They were asked to choose which of the six alternatives the synthetic items sounded most like.

The main issue here is to assess first of all which of the range of voices listeners find the most acceptable; and secondly, to relate this to how the phoneme recognition in the MRT test is affected by the use of different voices (male and female) and different voice types. It may be that subjects do not like the two male voices, therefore, are regarded as more acceptable than the female voice. There is still some way to go before a female voice can be automatically generated in this way, although in the earlier study [4], the female voice generated from the male copy-synthesis parameters gave the same naturalness ratings as the original male voice. However, the second male voice (V3), which is generated by transformation of the default voice, is preferred to that voice: although the target values for the parameters in the synthetic speech are based on a real speaker, that voice is not perceived to be the most acceptable.

All main effects and two-way interactions were significant at the 1 per cent level; only the three-way interaction subject × utterance × type proved significant (at the 5 per cent level only). Some 16.6 per cent of the variance is attributable to differences in the mean ratings of each of the three subjects; since individuals will offer different in the range of ratings they use in such a test, we shall ignore these differences for the time being. By far the greatest effects seem to be from voice (accounting for 14.5 per cent of the variance) and type (39.3 per cent of the variance in this analysis). Since voice and type are the most important issues here, we shall concentrate on these effects and their interaction in greater detail.

Two-way interactions are significant for the mean ratings. V3 is always perceived to be more acceptable than V1 and V2, in line with the speech of the default settings suggested by others. The pre-emphasised speech (that is with the lip-radiation filter only) yields the poorest acceptability ratings.

Let us now consider the interaction between voice and type.

Again, a Tukey's HSD reveals significant differences for the mean ratings. V3 is always perceived to be more acceptable than V1 and V2 in conjunction with types T1, T3. In addition, V1 is only more acceptable than the female voice V2 when generated with T3; with T1 settings (the optimal values from [4]), the differences disappear. The mean acceptability ratings here in the acceptability of the female by comparison to the male voice are therefore due to the type. These differences are significant at the 1 per cent level.
with T1 is in keeping with the findings of [4]. Furthermore, no differences in acceptability are found between any of the voices with T2; this type is clearly of little use for acceptable synthetic speech.

With respect to type, T1 always gives higher acceptability ratings than T2 or T3; additionally, for V1 and V3, the male voices, only T3 settings yield significantly better ratings than T2. The optimal settings from [4] would still seem to be good for generating highly acceptable synthetic speech; straight pre-emphasis, which may conceivably help some phonemic contrasts though this remains to be seen, results in poor acceptability ratings.

Finally, the best mean acceptability rating for the synthetically generated speech (3.73 for V3 with T1) compares favourably with the naturalness ratings obtained for copy-synthesis of real speech in [4] (namely, around 3.33). This is encouraging for our text-to-speech system.

**MRT Test**

The analysis of this test is also continuing. Based on the results for four subjects only, however, a three-way fixed effects ANOVA (subject x voice x type) revealed no significant main effects or interactions. It is perhaps unwise to draw general conclusions at this stage. Given that proviso, however, we may assume that there will be little correlation between acceptability from the first listening test and recognition rate in the MRT. Voice and type apparently have little effect on the MRT score. As further results become available, we shall be in a better position to explore this further.

**PRELIMINARY CONCLUSIONS**

Listening tests have been run to assess the acceptability of a range of voice types, and the effect on intelligibility of those voice types for a text-to-speech system. Preliminary analyses of some of the results indicate the global settings and voices which produce the highest acceptability scores: a derived male voice is apparently the preferred one, though it appears that with the correct global parameter specification, an acceptable female voice might also be generated by the transformation of the default male voice rather than from a completely new set of parameters extracted from the speech of a real female speaker. At all events, these different voice types did not result, apparently, in different phoneme recognition rates in an MRT, though as more results become available this may in fact change. It cannot be ruled out, furthermore, that as the overall intelligibility of the system improves, listeners may become more sensitive to effects associated with different voices and types. At this stage, however, we can continue to develop a multi-voice text-to-speech system with high subjective acceptability ratings and without detriment to intelligibility.

**REFERENCES**