# FACTORS AFFECTING PERCEIVED QUALITY AND INTELLIGIBILITY IN THE CHATR CONCATENATIVE SPEECH SYNTHESISER

*Nick Campbell, Yoshiharu Itoh, Wen Ding, and Norio Higuchi*

ATR Interpreting Telecommunications Research Laboratories
e-mail: nick@itl.atr.co.jp, http://www.itl.atr.co.jp/chatr

## ABSTRACT

In order to eliminate trial-and-error in the process of selecting a good speech database as a voice source for concatenative speech synthesis, and to determine the acoustic and prosodic characteristics that best predict 'appeal' or perceived 'quality' in the synthesised speech, we performed tests to evaluate listener preferences over a range of different synthesised voices. We found that variation of fundamental frequency in the source database, and open-quotient of the glottis as measured by joint-estimation (ARX) were the best correlates. To our surprise, there was very little correlation between the scores for 'intelligibility' and those for 'naturalness' in the test data, but the former showed a close correlation with durational characteristics, and the latter with pitch and loudness.

## 1. INTRODUCTION

Speech synthesis has undergone considerable improvements in voice quality with the development of concatenative systems which use recordings of human speech as the source material for waveform generation. It is now possible to generate synthetic speech in the voice of a person that is instantly recognizable as that of the original speaker, although the intonation and speaking style may not yet faithfully reproduce the characteristics of natural human speech.

While the majority of concatenative speech synthesis systems still make use of signal processing techniques to modify the prosody of the generated waveform, the CHATR system [2, 3], currently being developed at ATR, eliminates this stage by replacing it instead with prosody-based selection of phoneme-sized waveform segments from a large corpus of naturally-produced speech. The benefit of our approach is that it allows simple concatenation of very high-quality recordings and thereby maintains all the variation and fine details of the original speech. The damage done to the naturalness of the recordings by stretching and warping their original prosody is eliminated.

A consequence of this approach is that the speech corpus can be considered external to the synthesiser, which becomes instead an indexing device that provides a set of pointers into a sequence of segments from different locations in the original speech, which will join together smoothly to form a novel utterance. The process of creating the original index is time- and cpu-consuming, but need only be performed once for any given corpus.

We have produced more than thirty voices to date, for four languages, typically using about 40 minutes of speech from each, though the amount needed for good quality synthesis appears to vary according to language (less for languages with fewer vowels) and speaking style of the original recordings (increasing with the spontaneity of the speech). Different speakers, or speech corpora, are received differently by different listeners to the synthetic speech, and in judging the quality of the synthesis there can easily be confusion with personal preferences in matters of voice-quality.

## 2. MATERIALS

In an attempt to quantify the acoustic properties of different voices and relate them to 'popularity' in terms of 'likeable' speech synthesis, we performed an evaluation of perceived quality over several dimensions using voices synthesised from the waveform databases of 15 speakers, seven male and seven female Japanese speakers and one Japanese-speaking English male.

We asked 15 listeners (all Japanese) to grade the quality of 10 sentences of speech synthesised using the 15 different speech databases and compared the ranking of the results to acoustic and prosodic characteristics of the speech waveforms in each corpus.

The 10 test items were selected at random from a phoneme-balanced set of 503 magazine and newspaper sentences, and synthesised with the text-to-speech component of CHATR using the fifteen different speakers' speech waveform databases.

The resulting 150 sentences were presented to the subjects five times, in different randomised sequences, over a period of two weeks. The subjects were asked on five separate occasions (a – e) to evaluate each of the 150 sentences in terms of a) 'over-

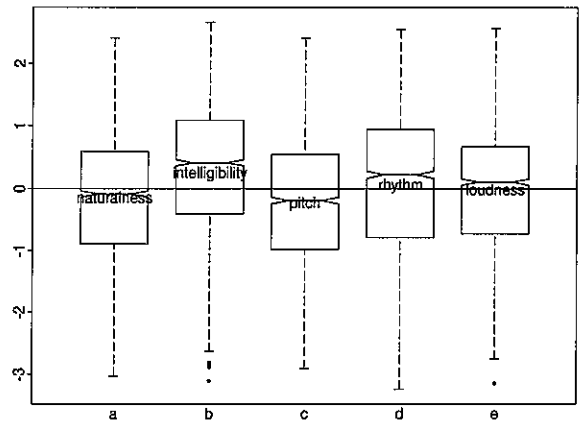


Figure 1: Normalised scores (zero mean) per test

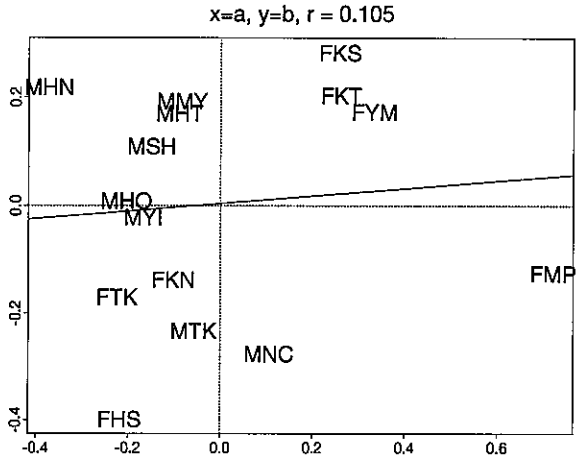


Figure 2: Almost no correlation between normalised scores for 'naturalness' (a) and 'intelligibility' (b).

all naturalness', b) 'ease of intelligibility', c) 'appropriate intonation', d) 'appropriate rhythm', and e) 'appropriate loudness'. They were given identical score sheets each session, offering choices ranging from 'good' to 'bad' in five steps. Five dummy sentences were included at the beginning of each session to allow subjects a practice period. The scores from these five extra sentences were not included in the analysis.

The purpose of the experiment was not to obtain data specific to each analysis dimension, although duration, pitch, and power are variables under synthesiser control, but rather to form a ranking of the averaged scores so as to grade the fifteen speakers in order of overall listener preference, and thereby to enable an analysis of the acoustic and prosodic characteristics of the more popular voices. If we can predict the 'quality' or 'appeal' of the synthesised speech from measurable characteristics of the original recordings, then we can save a lot of trial-and-error in the speaker-selection and database-selection aspects of voice synthesis.

## 3. TRANSFORMS

Some voices were liked more than others, some listeners were more severe in their scoring than others, and some sentences were scored higher than others. Analysis of variance of the resulting scores showed significant effects for speaker, but also effects for test item, test type, and listener. The raw results were therefore z-score normalised to remove subject, sentence, and test dependencies (by subtracting the mean and dividing by the standard deviation to produce a unitless value in the range of $\pm 3$) after factoring for each of the above variables. This transform produced a set of scores that were listener independent and item independent, to allow measures from the different tests to be generalised and compared at the same level.

## 4. RESULTS

Raw scores for all five tests (a – e) tended to show a positive bias. On a scale of 0 (bad) to 4 (good) a neutral score would be 2.0, and our evaluation averages were a=2.11, b=2.61, c=2.00, d=2.37, and e=2.27. Scores for 'intonation' were lowest[1], and 'intelligibility' (test b) scored highest of all (see figure 1). After normalization, significant effects were still found for the variable 'speaker' ($F_{(14,11235)} = 45.10$). Subjects performing the ranking were balanced for sex: 8 males and 7 females, but there was an effect found for sex of the speaker. Differences in preference for female voices ($F_{(6,5243)} = 96.84$) were more marked than those for male voices ($F_{(7,15992)} = 5.75$).

We cannot be sure what listeners were paying attention to (or being otherwise influenced by) when scoring each test under the different categories, but there are interesting correlations between the general category scores (a,b) and the more specific scores (c,d,e). As table 1 shows, there is a very strong correlation between scores for (c) and (e), but there was little evidence of a relationship between (a) and (b) (figure 2). Contrast this finding with the clear correlations shown in figure 3. We had assumed that subjects would find most 'natural' ($a$) the sentences that they found most 'intelligible' ($b$), but this appears not to be the case. Sentences scored highest for intonation were considered most natural, but those scored highest for 'rhythm' were on the other hand considered more intelligible. Further research is still necessary in this area.

Because of this complex interaction between the scores, we performed a Principal Component Analysis [1] to determine the main axes of discrimination,

[1] See [5] for further discussion of recent $F_0$ prediction and selection improvements in CHATR

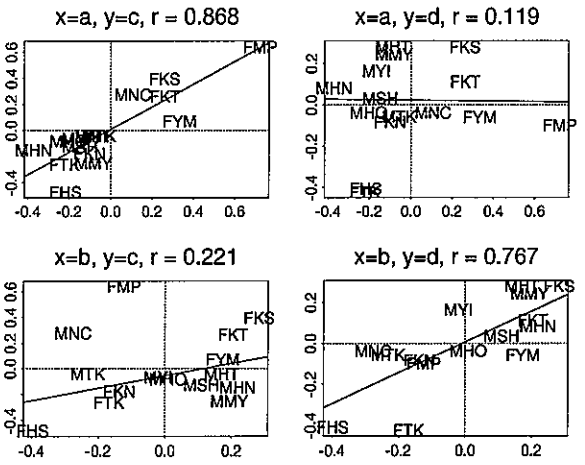


Figure 3: Interesting correlations between 'naturalness' (a), 'intelligibility' (b), 'pitch' (c), and 'duration' (d).

Table 1: Correlations among preference scores

| (a) naturalness | (a) | (b) | (c) | (d) |
|---|---|---|---|---|
| (b) intelligibility | 0.105 | — | — | — |
| (c) pitch | 0.85 | 0.129 | — | — |
| (d) duration | 0.12 | 0.768 | 0.39 | — |
| (e) power | 0.86 | 0.221 | 0.92 | 0.34 |

and computed a ranking based on the weightings determined for each factor (see figures 4 and 5).

Comparing the ranked preference score with acoustic and prosodic features of the various speech databases, we found strong correlations with both $F_0$ and glottal characteristics (estimated by an ARX model [4]). Figure 7 shows a plot of ranked scores against mean $F_0$, and against the standard deviation of the fundamental frequency for each sentence synthesised. Preference scores for both sexes tended to be higher for a lower $F_0$. For both sexes, there was a clear preference for speech generated from a database with less variation in $F_0$. In predicting the preference scores, we find standard deviation of $F_0$ in the raw speech database to produce the best correlation, with $r = 0.64$ for males and $r = 0.66$ for female voices.

Performing a linear regression on the ranking of scores using the glottal components from an ARX analysis [4] as factors, we can account for 50% of the variance by glottal parameters alone, and find that of these, the open quotient accounts for the largest part (r=0.93). The first two components account for 77% of the loading (figure 6), and including the third reaches 99%.

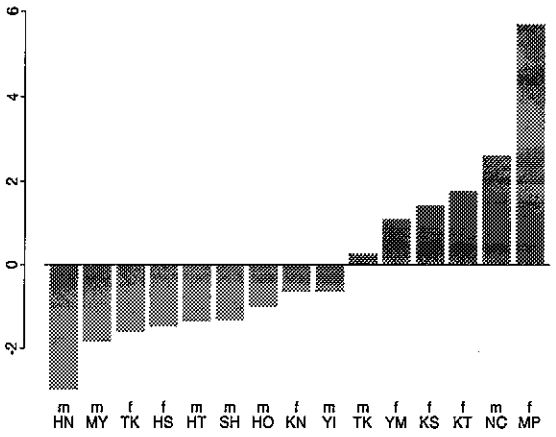


Figure 4: Ranking the speakers by summing scores on the first two principal components

[A0193MHN.WAV A0193MMY.WAV A0193FTK.WAV A0193FHS.WAV A0193MHT.WAV A0193MSH.WAV A0193MHO.WAV A0193FKN.WAV A0193MYI.WAV A0193MTK.WAV A0193FYM.WAV A0193FKS.WAV A0193FKT.WAV A0193MNC.WAV A0193FMP.WAV]

Table 2: Correlations among prediction factors (f0=fundamental frequency, oq=open quotient, gn=glottal noise, and st=spectral tilt)

| | intercept | (f0) | (oq) | (gn) |
|---|---|---|---|---|
| f0 | 0.4535 | — | — | — |
| oq | -0.9284 | -0.7004 | — | — |
| gn | -0.5795 | -0.4108 | 0.5372 | — |
| st | 0.3096 | 0.6964 | -0.6183 | -0.2517 |

## 5. DISCUSSION

There is a clear correlation between the variation in $F_0$ of a given speech corpus and the perceived quality of the synthesised speech when generated by concatenation of phoneme-sized waveform segments without subsequent signal processing. However, we found that judgements of good intonation in the synthesis had very little correlation with perceived intelligibility of the utterance, and that judgements of good duration or rhythm were more relevant.

In general, female voices seem to be preferred over male voices, and the English male voice (speaking Japanese) stood out from other males in this evaluation. We suspect that listeners may have been more sympathetic to the speech of a 'foreigner', though many subjects, when asked afterwards, claim not to have noticed the difference.

Two speakers (FHS and FTK) stood out as markedly distinct. Both had been recorded using a head-mounted microphone, and we assume that its 'noise-cancelation' effect (which cuts low-frequency spectral energy and boosts higher frequencies) may

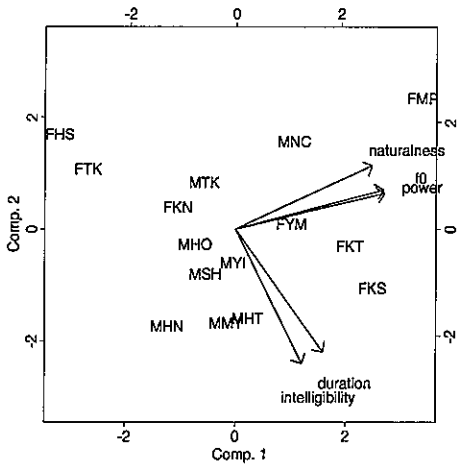


Figure 5: First two principal components account for 91% of the scores in two orthogonal dimensions

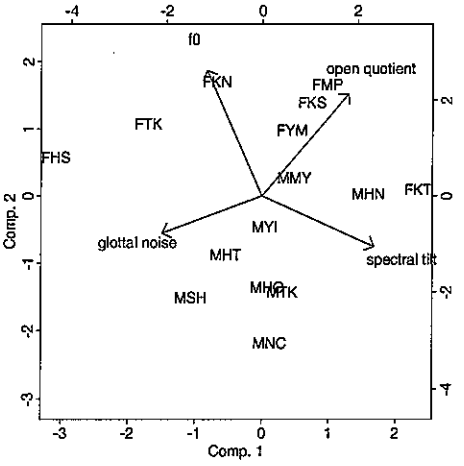


Figure 6: PCA distribution by Glottal components

be the cause. Integration of the signal using a low-pass filter can perhaps be applied as a post-process to reduce this spectral imbalance.

Good prediction of preference scores was achieved from analysis of the glottal components, which confirms that the speaking style of the original recordings is an important factor. The open quotient determines not only the amount of power in the speech waveform, but also the softness of the voice, the timbre of the speech sound. Titze [6] describes it as controlling the range between 'brassy' and 'fluty' voice. This may reflect the speaker's state of relaxation. Experience has shown that rather than have readers struggle to produce unnatural 'phonemically balanced' sentences for the sake of a representative speech database, it is better to have them talking for longer but in a more relaxed way and to obtain the balance through the collection of more speech data.

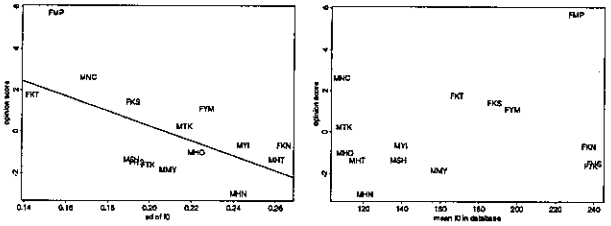

Figure 7: predicting scores from database mean F0 (left) and SD of the database F0 (right)

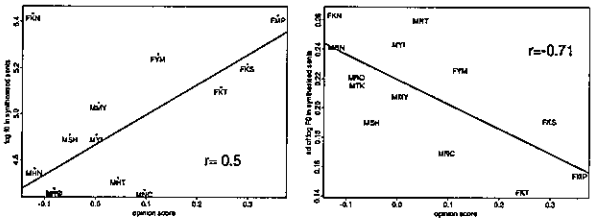


Figure 8: high mean and small SD preferred in the synthesised utterances

## 6. CONCLUSION

We have evaluated listeners' preferences for a range of different synthesised voices and found the strongest predictor of 'quality' in the synthesised speech to be variation of fundamental frequency in the original recordings. Tension of the original voice was the other main predicting factor. By concentrating on intonation, we can improve naturalness. However, to maximise intelligibility, we need to improve not $F_0$, but duration. Rather than do this by signal processing, which degrades the quality of the speech, we prefer to increase the size of the source corpus to include more natural variation. By understanding the dimensions which most influence listeners perceptions, we can select more appropriate source corpora.

## REFERENCES

[1] Becker, R. A., Chambers, J. M. & Wilks, R. A. **"The New S Language: A Programming Environment for Data Analysis and Graphics"**, Wadsworth & Brooks/Cole, California (1988).

[2] W. N. Campbell & A. W. Black, "CHATR: a multilingual speech re-sequencing synthesis system", 45-52, SP96-7 Tech Rept IEICE, 1996(5).

[3] W. N. Campbell, " CHATR: A High-Definition Speech Re-Sequencing System", Proc 3rd ASA/ASJ Joint Meeting, 1223-1228, Hawaii, 1996(12).

[4] W. Ding, H. Kasuya, and S. Adachi, "Simultaneous estimation of vocal tract and voice source parameters based on an ARX model", *IEICE Trans. Inf. & Syst.*, Vol. E78-D, pp. 738-743 (1995)

[5] K. Fujisawa, T. Hirai, N. Higuchi, "Use of pitch-pattern improvement in the Chatr speech re-sequencing system", this volume.

[6] I. Titze, **"Principles of Voice Production"**, Prentice Hall, N.J., (1994).