# INVESTIGATION OF ANALYSIS AND SYNTHESIS PARAMETERS OF STRAIGHT BY SUBJECTIVE EVALUATION

*Parham Zolfaghari*** *Yoshinori Atake*† *Kiyohiro Shikano*† *Hideki Kawahara***

*CIAIR/CREST, Itakura Laboratory, Nagoya University, Furo-cho 1, Chikusa-ku, Nagoya 464-8603, Japan.
**email :** [illegible]
†NAIST, Ikoma, Nara, Japan
**Wakayama University/CREST/ATR Human Information Processing, Wakayama, 640-8510 Japan

## ABSTRACT

The goal of this paper is to locate and understand the fine fundamental problems that exist in the representation of speech sounds by a very high quality speech analysis/synthesis engine namely STRAIGHT. The approach followed here is the evaluation of this system using subjective measures. We use the diagnostic rhyme test (DRT) to evaluate the intelligibility of speech analysed and synthesised by this system for various analysis frame-rates. Consequently we catagorise the fine problems and suggest possible improvements. The results from the DRT have indicated that STRAIGHT can produce speech with an average DRT score of 95 between 1-5 ms analysis frame-rate. In addition, a set of subjective quality measures using MOS and MNRU tests have been conducted. These tests have been carried out for three different versions of the STRAIGHT system: versions 17, 23 and 30. The DRT has been carried out using version 23 only. Based on the subjective evaluation results, a discussion of possible improvements to the STRAIGHT system is given.

## 1. INTRODUCTION

A system capable of reproducing very high quality speech sound while enabling flexible control on physical parameters of speech would be an important tool for speech perception research as well as other speech research applications. The aim of this study is to investigate fine details where the STRAIGHT speech analysis/modification/synthesis system [2, 5, 4] fails to replicate important speech attributes. This system is a research tool used by a number of institutions and laboratories. It is known to produce speech of very high quality.

This paper will present subjective listening test results for the STRAIGHT system. Subjective tests can be divided into two categories: intelligibility testing and quality testing. The two classes are not disjoint, and good quality implies good intelligibility while the converse is not necessarily true. The intelligibility test used in this study is the diagnostic rhyme test (DRT) [7] on STRAIGHT operating at various analysis frame-rates. This is then followed by perceptual quality evaluations of this system using the Mean Opinion Score (MOS) and Modulated Noise Reference Unit (MNRU - ITU-standard P.810) tests in clean and noisy speech environments.

Three versions of STRAIGHT (versions 17 [2], 23 [5], and 30 [4]) have been used in the quality tests and only version 23 was used in the intelligibility tests. Version 30 was developed from the results of the version 23 DRT scores. In the following sections, after reporting on the evaluation results, a discussion of the relationship among the parameters in the STRAIGHT versions in terms of quality and intelligibility measures will be given.

## 2. THE STRAIGHT SYSTEMS

The source-filter model has been shown to produce synthetic speech of very high quality [6]. However, for such systems to attain this high quality the parameters need careful attention and hand-editing. The STRAIGHT system, based on the source filter model, allows flexible control of speech parameters and its conceptual simplicity has made this system a tool for speech perception research as well as other speech research applications. This has been achieved by refining the procedures for extracting the fundamental frequency (F0) and spectral envelope, and a procedure to generate sophisticated excitation signals in synthesis. Four versions of the STRAIGHT system that have been developed are as follows: versions 14, 17, 23, and 30. Many researchers and developers using this system have their own preferred versions of the system which they believe produce the best quality speech. The following sections will lead to a discussion on which of these versions is perceptually better and what parts of the system can be improved. In general, it is instructive to investigate closely on deficiencies for finding limitations of an underlying model. In this case, a source filter model and implicitly, principles of Auditory Scene Analysis [1].

The upgrade from version 17 to 23 was designed to make STRAIGHT more accessible for researchers without having strong DSP knowledge. The only major difference is in the group delay parameter setting for synthesis. A group delay parameter is used to design each of the excitation pulses. The group delay function of one such pulse is generated from a spatial-frequency-band-limited random signal. The standard deviation of the group delay function in a high frequency region is set according to the group delay parameter. In STRAIGHT version 17, this standard deviation is set to $0.4x1000/F0$ ms using this group delay parameter. In other STRAIGHT versions, a fixed 2 ms standard deviation is set by default. The standard deviation in the lower frequency region is suppressed using a sigmoidal function.

Version 30 was designed based on the DRT evaluation finding to be reported in the following section. Also, a new fundamental frequency extraction scheme based on frequency to instantaneous frequency mapping is implemented in this version [4]. In the other two versions, F0 is extracted using an instantaneous frequency method with an adaptive fundamental component selection mechanism [5].

## 3. DRT INTELLIGIBILITY EVALUATION

The DRT tests the ability of listeners to distinguish phonemes with common attributes. The procedure for DRT requires one word of each rhyming group of words to be presented to listeners, and they are asked to pick the word that was spoken. The test material consists of 96 rhyming monosyllable word pairs that were selected to differ in only their initial consonant (e.g. veal-feel) [7]. Each word pair tests for distinctive features including *voicing*, *nasality*, *sustention*, *sibilation*, *graveness* and *compactness* and scores in each of these categories provide information on diagnosing system deficiencies. These word pairs were recorded by an English native speaker at 16 kHz in a recording room. Tests were carried out in a listening booth at ATR Human Information Processing Labs, Kyoto, Japan.

### 3.1. Analysis of DRT scores

The aim of this experiment is to see the difference in intelligibility of STRAIGHT after varying the analysis frame-rate. Frame-rates used were 1 ms, 3 ms, 5 ms, 7 ms, 10 ms, 16 ms and 20 ms. Six male native English listeners were used for the DRT. The rate for the presentation of DRT items was one item every 1.5 seconds. Only STRAIGHT version 23 was used in this evaluation.

The DRT scores obtained for the STRAIGHT system are shown in Figure 1. These results demonstrate that STRAIGHT is a highly intelligible speech sound representation system. The speech analysed with STRAIGHT using frame-rates in the region of 1 to 5 ms obtained similar and very high DRT scores and a steady log-linear decrease in intelligibility can be seen after this region. Benchmark tests on the original speech at 16 kHz obtained a DRT score of 99.1.

In the lower frame-rates of 1 to 5 ms, categories of *voicing*, *sustention*, and *graveness* show lower intelligibility. The reduction in discriminability of the *voicing* category can be attributed to system deficiencies with respect to other acoustical-speech features which are perceptually significant; in particular the time of onset of periodicity. In our test for the *voicing* category, the listeners had difficulty in distinguishing between phonemes /d/ and /t/ as in DUNE and TUNE as well as DINT and TINT. GIN and CHIN also showed poor discriminability. Referring to spectrograms of Figure 2 (DUNE) and Figure 3 (TUNE), in the case of TUNE we can see a stronger and longer segment unvoiced initial consonant. All listeners who made a mistake on this stimuli word pair mistook DUNE as TUNE. This suggests that the voicing decision in STRAIGHT is not reliable enough in the such unvoiced segments which in effect increases the voice onset time.

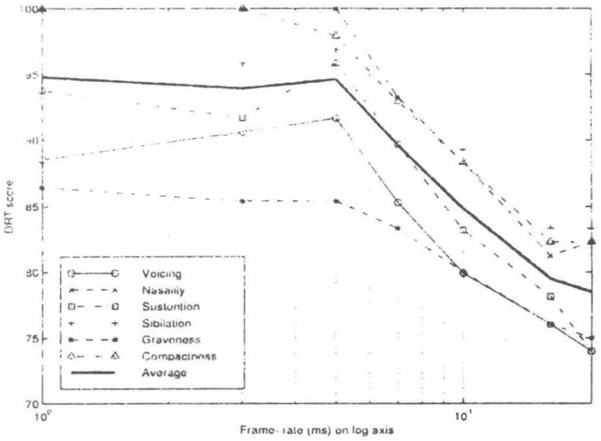

Figure 1. DRT scores for different analysis frame-rates.

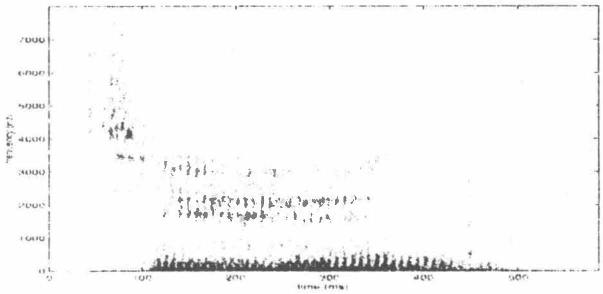

Figure 2. Wide-band Spectrogram of stimulus DUNE.

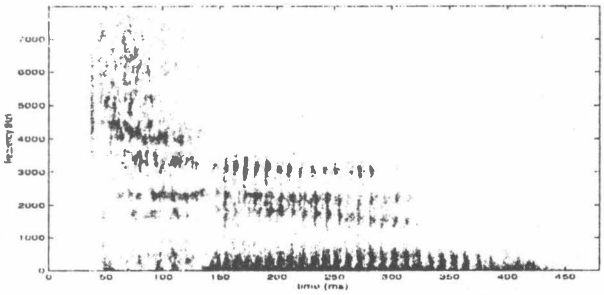

Figure 3. Wide-band spectrogram of stimulus TUNE.

In the *sustention* category which distinguishes between items consisting of sustained and interrupted word pairs, listeners had over all analysis frame-rates, problems with SHOES and CHOOSE, VON and BON, VOX and BOX, DOZE and THOSE, and DAN and THAN. In this category, it is important on fine temporal modelling of the first consonant otherwise the word pairs are spectrally very similar.

In the case of *graveness*, distinction between phonemes /f/ and /Θ/ was found difficult. These include word pairs such as FIN and THIN, FOUGHT and THOUGHT and FAD and THAD. Similar to some of the stimuli in the sustension category, this is mainly due to the type of windowing and smoothing in the time domain which is a poor representation of abrupt bursts such as in the word pairs FAD and THAD.

Furthermore, KEEP and CHEEP in the *sibilation* category produced a number of errors by listeners at frame-rates above 1 ms. In this category, modelling noise in the consonant region plays an important part since there is a difference in the intensity of this first consonant in the frequency domain, though the duration of the consonants are similar.

## 4. QUALITY EVALUATIONS

Objective quality evaluation was carried out using MOS and MNRU tests. Three versions of STRAIGHT (versions 17, 23, and 30) were subjected to these tests.

### 4.1. Test Conditions

These tests were carried out at NAIST's specially designed listening rooms. The listeners included fifteen students between the ages of 22 and 24 years, two of whom were female. The sounds presented to the listeners are listed as follows:

1. Original utterances;
2. Utterances analysed and synthesised by the three versions of the STRAIGHT system;
3. Utterances coded by G.722(SB-ADPCM) 64kbit/s;
4. Noise correlated to the amplitude of utterances (MNRU equivalent-Q).

These utterances were sampled at 16 kHz and the speakers were three males and three females each uttering 6 sentences in Japanese. Additive noise was of type white and pink using SNR of 10 dB, 15 dB, and 20 dB.

### 4.2. Evaluation results

Figure 4 shows the MOS and MNRU results obtained for the different versions of STRAIGHT using clean speech. It is observed that version 17 is of the highest quality which may indicate that fundamental frequency adaptive group delay representation in synthesis (refer to Section 2) results in higher quality. It is also noted that male speakers show higher objective scores. This is mainly due to the difficulty in extraction of spectral information for female voice since the higher fundamental frequency of females results in a more complex spectrum. Overall, the STRAIGHT system scores lower in quality in all versions to that of the original voice, which is expected, and that of G.722 codec.

Figure 5 shows the quality evaluation scores obtained after addition of white noise to the signals. Version 23 of STRAIGHT seems to be more robust to noise. However, in all versions the scores are quiet low and not significantly different from each other. The ratio of quality difference between original and coded utterances is higher than that for clean speech. The main reason for this degradation is that the source information is highly influenced by white noise. More specifically, spectral information is highly corrupted.

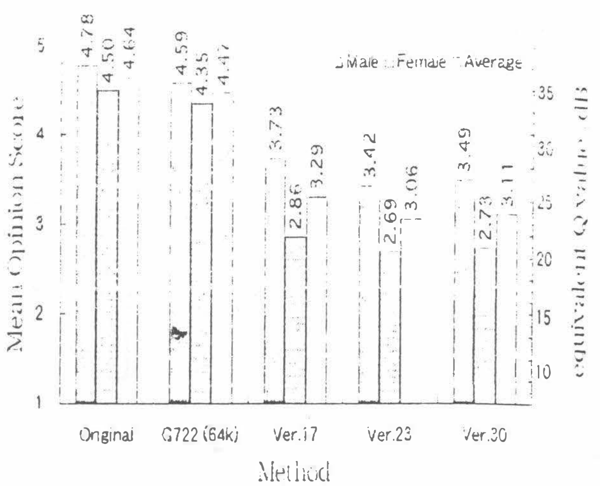

Figure 4. MOS and MNRU scores for clean speech.

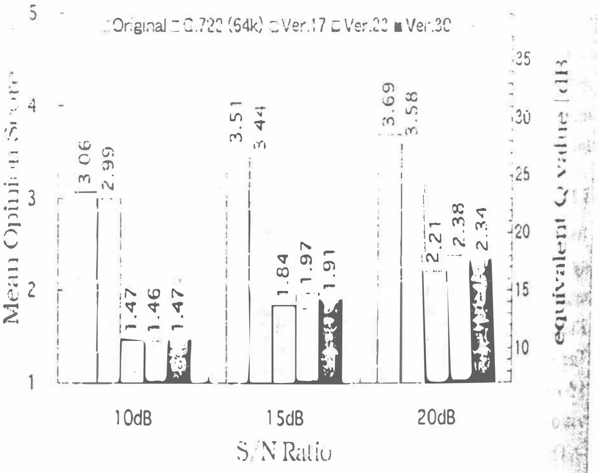

Figure 5. MOS and MNRU scores for speech with added white noise.

The quality evaluation scores after the addition of pink noise to the utterances are shown in Figure 6. STRAIGHT version 23 performs slightly better than the other versions. This is no significant increase in the quality. Degradation effect in the source information is quite high.

## 5. DISCUSSION

It is to our surprise that version 30 did not perform better than the other versions. As mentioned, version 30 was revised based on the analysis of DRT evaluation results for STRAIGHT version 23. However, there is a promising view point on this matter discussed in this section, and it is evident that these subjective listening tests have clarified STRAIGHT's deficiencies and close investigations on these suggest further methods to refine STRAIGHT. It is also indicated that, at least up to this point, the average DRT score of 95 does not imply the basic limitation of the source-filter model.

In the objective evaluation tests, listeners reported the quality

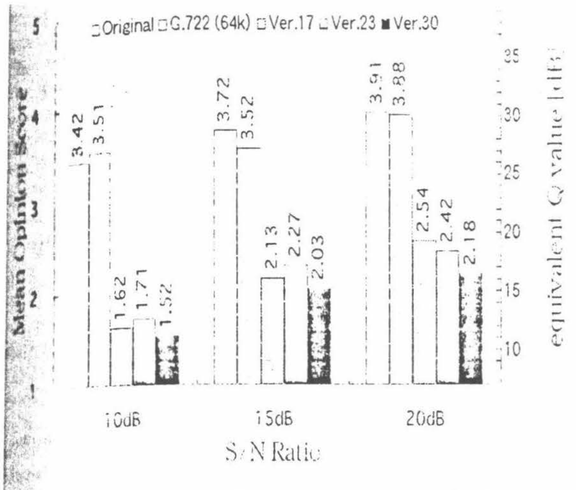

Figure 6. MOS and MNRU scores for speech with added pink noise.

of some voiceless sounds to have a buzzy characteristic associated with them such as in the phoneme /ʃ/. This was specifically apparent for version 30 of STRAIGHT. Since the fixed points method [4] in this version is very effective in picking any (fundamental-like) periodic components, it is possible to pick up periodic components such as AC-current noise and fan noise from air conditioners. If such erroneous selection occurs in silence regions or in voiceless fricative regions, higher frequency components may be driven by a periodic pulse (due to a voiced decision in an unvoiced region) and produce sound of buzzy characteristic.

Two methods of solving this problem exist and these can be incorporated collectively. One is to introduce a decision rule to identify if the extracted periodic component is due to speech or noise. The other is introduce a type of multi-band excitation (MBE) scheme, where each frequency band is excited by a mixture of periodic and aperiodic driving sources. Although the current version of STRAIGHT employs such MBE representations, when the quality tests were being performed, version 30 did not have the MBE type source control function. This was due to the large number of parameters in the MBE implementation which were not fully optimized at the time of the tests.

Moreover, some side effects still exist from the modifications made to fix the problems in *sustention* and *sibilation* categories of the DRT evaluation. The deficiencies in these categories indicated that frequency resolution for unvoiced sound analysis is not good enough. To overcome this the time window length for unvoiced sound analysis was increased. The poor temporal resolution due to this long time window is resolved by multiplying the short term power envelope with the smoothed spectrogram produced by STRAIGHT's spectral smoothing method [2]. This temporal shaping is also intended to solve the DRT problem in *sibilation* and *graveness* categories due to the poor representation of abrupt bursts and stop consonants. However, the implementation of this temporal shaping is too complex and ad-hoc that no formal mathematical analysis is possible. As a consequence, this high frequency resolution combined with a fine temporal envelope, resulted in unvoiced sounds having a musical noise or "wet" color characteristic to them. Simple MBE type source control does not solve this problem. It is necessary to adaptively change the time-frequency smoothing scheme, based on periodicity and sharpness of an event in the source speech signal. A novel event detection method [3] provides an indispensable information to make the above decisions reliable and mathematically well defined.

In order to compete with G.722(SB-ADPCM), tolerance to noise of the STRAIGHT systems requires improvements by the use and adaptation of some speech-in-noise models. The spectral envelope smoothing and fundamental frequency extraction of this system require particular attention in these noisy conditions.

## 6. CONCLUSIONS

A report on the results from subjective evaluation of STRAIGHT using the DRT, MOS and MNRU tests have been given. Although these results are for this specific system, our intention is also to demonstrate the areas where a speech analysis/synthesis system may need to pay attention in order to produce synthesised speech of very high quality. The results from this study revealed that the DRT score of the current implementation of STRAIGHT saturates at frame-rates shorter than 5 ms to an average of 95. The objective evaluations using MOS and MNRU tests revealed good quality scores across the three versions of STRAIGHT and white and pink noise tolerances were reported.

## REFERENCES

[1] BREGMAN, A. S. *Auditory Scene Analysis: Perceptual Organization of Sound.* Bradford Books, 1994.

[2] KAWAHARA, H. Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisted. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (Munich, April 1997), vol. 2, pp. 1303–1306.

[3] KAWAHARA, H. Accurate vocal event detection method based on a fixed-point analysis of mapping from time to weighted average group delay. In *Proceedings of ICSLP* (2000). to appear in.

[4] KAWAHARA, H., KATAYOSE, H., CHEVEIGNÉ, A. D., AND PATTERSON, R. Fixed points analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity. In *Proceedings of EUROSPEECH'99* (1999), vol. 6, pp. 2781–2784.

[5] KAWAHARA, H., MASUDA-KATSUSE, I., AND DE CHEVEIGNÉ, A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction. *Speech Communication 27*, 3-4 (1999), 187–207.

[6] KLATT, D. H. Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America 82*, 3 (September 1987), 737–793.

[7] VOIERS, W. Evaluating processed speech using the Diagnostic Rhyme Test. *Speech Technology 1* (1983), 338–352.