ASSESSMENT OF THE DEGRADATION OF SYNTHETIC SPEECH AND TIME FREQUENCY WARPING OVER DIFFERENT LISTENING LEVELS

M D Burrell

Aston University, Department of Electrical and Electronic Engineering and Applied Physics, Birmingham, England

1. INTRODUCTION

Variability of subjective scores can to a certain extent, be mitigated by the use of certain forms of "reference device". Various forms of reference devices have been used for assessing the transmission quality of telephone connections. For example transmission loss, injected circuit noise, interupted speech and speech correlated noise. Choice of reference device is governed by the form of the predominant degradation and the types of perceptual errors incurred. None of the reference devices previously used seem particularly suitable for use with synthesized speech now being introduced into public telephone networks. British Telecom labratories, Martlesham Heath, proposed a new reference degradation, Time Frequeny Warping (TFW), for assessing the synthesized speech. TFW is essentially modulation by well-defined and tightly controlled "wow" and "flutter". It has been shown that synthesized speech is less redundant than natural speech [1]. Hence, any degradation in speech quality introduced by a telephone connection, degrades the synthetic speech to a greater degree compared to natural speech. For TFW to be a viable reference device, it is desirable that subjectively equivalent TFW modulated and synthetic speech samples should degrade in a similiar manner under the same listening conditions. The paper presents results of some experiments which form part of a contract carried out for British Telecom. The work was to assess the suitability of TFW modulation of natural speech for use as a reference degradation for assessing the quality synthesized speech.

2. GENERAL CONSIDERATIONS

2.1 It is desirable for the speech produced by a reference device to be similiar to the speech it is being used to assess. Hence, an adjustable attenuator is used for assessing transmission loss, added speech interference, for sidetone and speech modulated noise is used for quantization effects. It may therefore be assumed that the resulting perceptual errors produced by the reference device and the test speech will be perceived as similiar. This approach enables the subjects taking part in subjective tests to use the same criteria for assessing the reference degraded speech and the speech being assessed. The in this study listening effort tests were used to assess the speech. Altough comparisons in terms of listening effort can be made between systems which sound very different, the listener`s task is made much easier if they do sound similiar. It cannot be claimed that TFW modulated real speech sounds exactly like any particular version of synthesized speech but it can be said the resulting displacements in timing do have some resemblence to errors produced by synthesized speech.

3. TIME FREQUENCY WARPED MODULATION (TFW)

3.1 TFW modulation is a combined phase and frequency modulation of speech, which can be thought of as a form of tightly controlled "wow" and "flutter" [2]. The effect is produced by sampling real speech, at a rate well above the Nyquist rate (8KHz), store it , and then reproduce the samples at a variable rate; with the mean rate equal to the original sampling rate, varying over a range extending equally either side of it. The effect of TFW modulation on a sinewave is shown below.

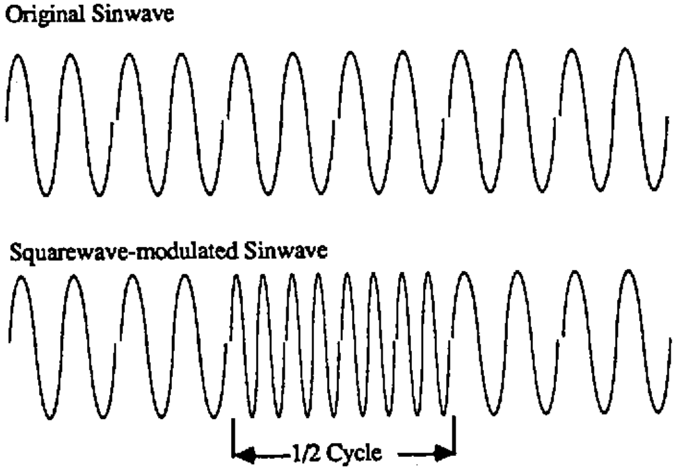


Figure 1

TFW modulation has three variable parameters for varying the output sampling rate, modulation waveform, period and amplitude. For a reference device it is desirable to have just one variable parameter, this simplifies the administration of comparison tests and increases the accuracy of repeated tests. In this study the comparison tests used were listening effort tests, therefore it was required that the range of degradation in speech quality produced by the TFW modulation was sufficient to cover the full listening effort range.

4. EXPERIMENTS

4.1 The aim of the experiments was to determine which TFW modulation parameters was the most suitable variable parameter and determine the values of the fixed parameters. Also to make some assessments in terms of listening effort [3], using typical examples of synthesized speech.

4.1.1 A pair comparison experiment was carried out to compare the three TFW waveform settings. Each TFW modulation wave form sinusoidal, square and triangular were compared using the same range of period and amplitude settings to a synthetic speech sample. The results showed the shape of the modulation waveform did not seem important as non were significantly different (better) than the others. It was decided to use the sinusoidal waveform because it produced the widest range of results for the period and amplitude settings used. There also appeared to be only a narrow choice of modulation period; it therefore seemed likely that the amplitude of the modulation period would be the suitable variable parameter. A listening effort experiment was conducted to investigate the relationship between TFW control parameter settings and listening effort scores, in order to determine whether just one control parameter produce a sufficient range of degradation in the speech to cover the whole listening effort scale.

4.1.2 Listening opinion tests have been used extensively in telephone network planning to compare different telephone links in terms of "listening effort". Subjects listen to groups of sentences and give their opinion of the speech according to the five category scale shown below.

Opinions based on the effort required to understand the meanings of sentences.

A	Complete relaxation possible ; no effort required.
B	Attention necessary; no appreciable effort required.
C	Moderate effort required.
D	Considerable effort required
E	No meaning understood with any feasible effort.

The subjects results are scored 4,3,2,1, 0, respectively, the mean of these values for each treatment is called "mean opinion score". For the listening effort experiments the sentences were recorded on a real-to-real tape recorder, and replayed via a Intermediate Reference System (IRS.) which represents a standardized local area network telephone connection.The experiment is based on a randomized 11* 11 graeco-latin square, in which rows represent listeners, columns represent the order in which treatments (speech type) are presented, the letter / number combinations or cells represent a "run" in which a particular list and treatment combination is replayed to the listener. Within each run the listening level is varied over a number of predetermined values in random order, one level per group of sentences. The design simplifies the analysis of the results and reduces undesirable effects, due to subjects sentences and sequence. The speech material were short simple sentences having no contextual relationship. Speech material was supplied by British Telecom [3].

Speech Treatments Used;

R1 - Natural speech

T1 - TFW,Modulation: Sine; Period; 20 ms; Amplitude 1 KHz
T2 - TFW,Modulation: Sine; Period; 20 ms; Amplitude 2 KHz
T3 - TFW,Modulation: Sine; Period; 200 ms; Amplitude 3 KHz
T4 - TFW,Modulation: Sine; Period; 150 ms; Amplitude 1.5 KHz
T5 - TFW,Modulation: Sine; Period; 200 ms; Amplitude 1 KHz
T6 - TFW,Modulation: Sine; Period; 200 ms; Amplitude 2 KHz
T7 - TFW,Modulation: Sine; Period; 200 ms; Amplitude 3 KHz

S1 - P.D.P.11 Text input synthesiser, variable pitch contour
S2 - P.D.P.11 Text input synthesiser, fixed pitch contour
S3 - Phoneme based system Mac. Talk Phonetic corrections

First Repetition

	R1	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
Mean	3.7	1.04	0.98	1.65	2.92	2.2	1.32	1.69	2.67	0.8	0.22
Second Repetition											
Mean	3.83	1.38	1.51	1.84	3.65	2.64	1.42	2.13	3.2	1.05	0.27

A mathematical model was calculated from the mean of the 1st and 2nd repetition's results, using multiple regression analysis. The model relates the listening effort score Y to the time frequency warping modulation settings period and amplitude.

Mathematical Model

$$Y = 4.29 - (5.5\times10^{-3}\ (\text{period})) - (9.24\times10^{-4}\ (\text{amplitude}))$$

Figure 2 represents the model equation which relates TFW modulation settings of period and amplitude to listening effort score Y. The figure is based on a grid were the x-axis and y-axis represent the amplitude and period of the TFW modulation, respectively . The listening effort scale Y was drawn by plotting equivalent LE. scores on the grid. The five listening effort scores 0 - 4 which represent the listening opinion scale, are shown as the thick diagonal lines. In a similar manner, the synthetic speech treatment scores were plotted and are represented by the thin diagonal lines. The TFW period and amplitude settings used in the listening effort test were plotted on the grid as black dots. The experimental results were plotted as spiked circles on the listening effort scale which runs diagonally from the bottom right corner to the top left corner. The points were plotted on a line which passes through the model equivalent period and amplitude settings and runs 90 degrees to the equivalent listening effort score lines.

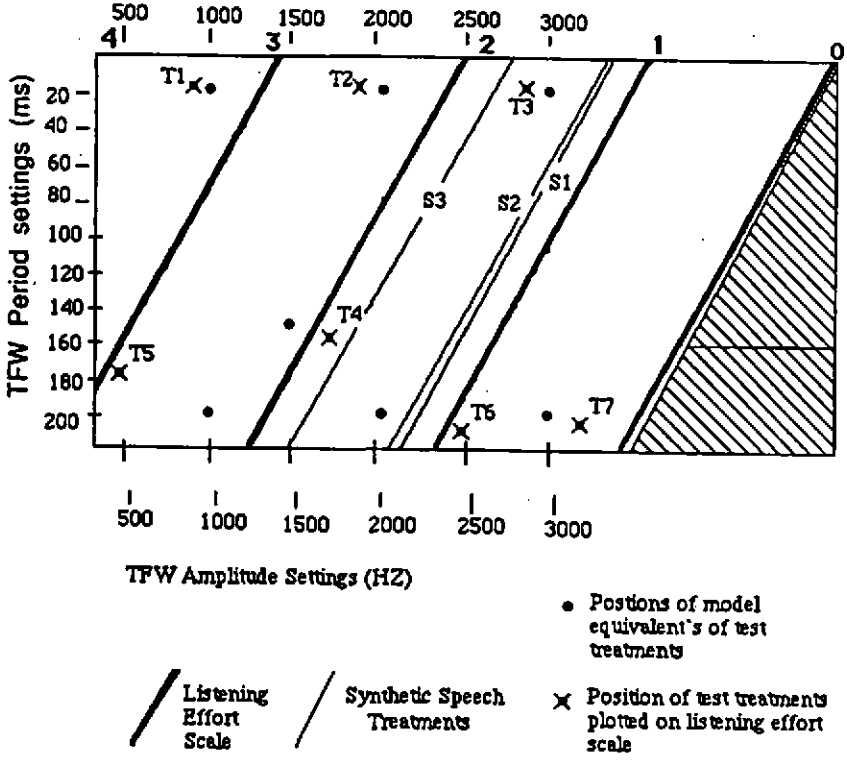


Figure 2

4.1.3 Selection of parameters; from figure 2, it can be seen that for a fixed period the range of listening effort scores that could be obtained by varying the amplitude setting is greater than the range obtained for a fixed amplitude and variable period. Hence, it was reasonable to adopt the variable amplitude at a fixed period as the TFW modulation variable parameter, a period of 150 ms was selected. From the above model, (fig 2) the highest LE. score for a period of 150 ms is 3.1 and the lowest is 0.6, which does not cover the upper values of the listening effort scale. This range can be increased by reducing the amplitude. If the amplitude is reduced to zero, the TFW modulation device would reproduce the original real speech, which, intuitively, would give the highest listening effort (Y), for the speech-sentence combination used. From figure 2 it can be seen that the synthetic speech treatments used, all had listening effort scores below 2, which indicates very poor quality speech. A listening effort score of 2.5 is taken as the point of on-set of listening effort, any speech with a score below 2.5 is unsuitable for use over a telephone network, because further degradation of the speech introduced by the telephone network would make the speech difficult or impossible to understand.

4.1.4 Further listening effort tests were conducted to determine the most suitable TFW amplitude settings and compare them with other synthesized speech treatments. The results of two tests are shown below:-
Listening Effort Test Jan

R1 - Natural speech
T1 - TFW,Modulation: Sine; Period; 150 ms; Amplitude 0.25 KHz
T2 - TFW,Modulation: Sine; Period; 150 ms; Amplitude 0.5 KHz
T3 - TFW,Modulation: Sine; Period; 150 ms; Amplitude 0.75 KHz
T4 - TFW,Modulation: Sine; Period; 150 ms; Amplitude 1 KHz
T5 - TFW,Modulation: Sine; Period; 150 ms; Amplitude 1.5 KHz
T6 - TFW,Modulation: Sine; Period; 150 ms; Amplitude 2 KHz
T7 - TFW,Modulation: Sine; Period; 150 ms; Amplitude 2.5 KHz
S1 - INFOVOX American version
S2 - INFOVOX 850905 "Preliminary British"
S3 - INFOVOX British (From Sweden)

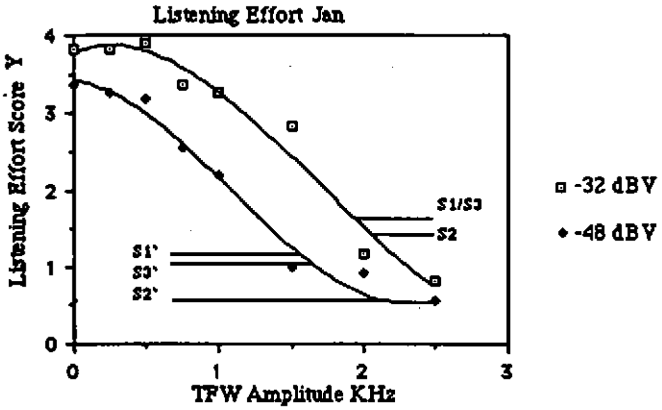
Listening Effort Test Jun

R1 - to - T7 same as listening effort test Jan
S1 - INFOVOX American version
S2 - INFOVOX 850905 "Preliminary British"
S3 - DEC Talk American

The mean subjective scores for two listening levels, -32 dBV and -48 dBV, relative to 1 volt across 600 Ohm attenuator, are shown below. The listening levels were measured using a SV6 speech volt meter ref ( ), which intergrates the speech voltage over a specified reference level, ignoring pauses of less than 10 ms.

Listening Effort Test Jan:- Listening Levels -32 dBV & -48 dBV

R1	T1	T2	T3	T4	T5	T6	T7	S1	S2	S3	
3.82	3.91	3.36	3.27	2.82	1.18	0.82	0.45	1.64	1.45	1.64	(-32 dBV)
3.36	3.27	3.18	2.55	2.18	1.00	0.91	0.55	1.18	0.55	1.09	(-48 dBV)



Listening Effort Jan

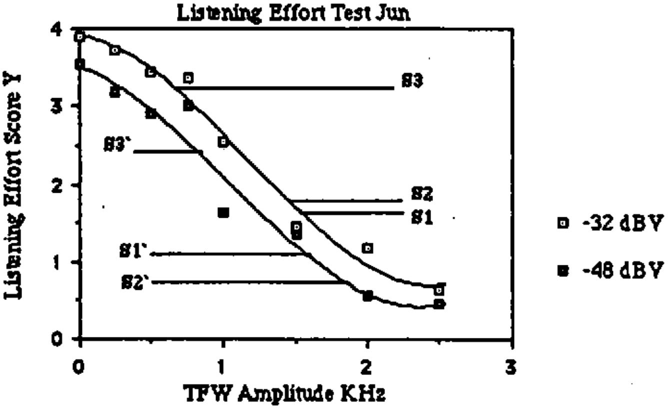
$y = 3.4176 - 0.2112x - 1.4975x^2 + 0.4545x^3$ R = 0.99

$y = 3.777 + 0.7507x - 1.5561x^2 + 0.3071x^3$ R = 0.98

Figure 3

Listening Effort Test Jun:- Listening Levels -32 dBV& -48 dBV

R1	T1	T2	T3	T4	T5	T6	T7	S1	S2	S3
3.91	3.73	3.45	3.36	2.55	1.45	1.18	0.45	1.64	1.82	3.27 (-32 dBV)
3.55	3.18	2.91	3.00	1.64	1.36	0.55	0.45	1.09	0.73	2.45 (-48 dBV)



Listening Effort Test Jun

$y = 3.5158 - 0.6526x - 1.1248x^2 + 0.3581x^3$ R = 0.98

$y = 3.9143 - 0.2312x - 1.4483x^2 + 0.411x^3$ R = 0.99

Figures 3 & 4 are graphs of the speech treatments plotted according to listening effort score and TFW amplitude. The real speech treatment was plotted as TFW = 0 and the synthetic speech treatments are plotted as lines corresponding to their listening effort scores. For the synthetic speech treatments in listening effort test Jun, a "equivalent TFW amplitude" (eTFW) was calculated. A "equivalent TFW amplitude" is a TFW amplitude setting calculated from the regression curve equations, which would produce a equivalent listening effort score to the synthetic speech treatment The results are shown below:-

listening Level -32 dBV	S1 - Y = 1.64 - TFW` equivalent 1.54 KHz
	S2 - Y = 1.82 - TFW` equivalent 1.43 KHz
	S3 - Y = 3.27 - TFW` equivalent 0.72 KHz
listening Level -48 dBV	S1` - Y = 1.09 - TFW` equivalent 1.57 KHz
	S2` - Y = 0.73 - TFW` equivalent 1.84 KHz
	S3` - Y = 2.45 - TFW` equivalent 0.8 KHz

The confidence limits for the "equivalent TFW amplitudes" depend on the regression curve confidence limits and the synthetic speech confidence limits, these two factors are not dealt with in this paper. The results given above, although not conclusive, provide an indication of the relationship between the degradation of a synthetic speech treatment and it`s "equivalent TFW amplitude" setting. It can be seen that the "eTFW" values for S1 at each listening level are very similiar, in fact the difference is not significant, this is seen for S3 as well. For S2 there is a greater difference in the "eTFW" values" for the two listening levels. If we consider figures 3 &4 it can be seen that at the extremes of the listening effort scale the difference between the listening levels, is in general less than it is in the middle of the scale. This maybe because subjects have less trouble in deciding on the effort required to understand the meaning of speech for high and low quality speech. Hence the variance of a sample mean for speech in the mid range of the listening effort scale maybe greater than at the extremes. If true, it would account for the differences in degradation of the various TFW modulated speech treatments.

5 CONCLUSIONS

5.1 The results indicate that subjects can compare time frequency warped and synthesized speech in terms of listening effort, which supports the results of the earlier work ref ( ). The results also provide evidence that reference equivalents can be established between time frequency warped speech and synthetic speech. This supports the proposal that time frequency warping should be used as a reference device for assessing synthetic speech.

Further analysis of the results of these and other listening effort tests are needed to clarify differences in the variance of subjective scores for different speech treatments.

6. REFERENCES

[1] H C NASBAUM, M J DEDNA & D B PISONI, 'Perceptual Confusions of Consonants in Natural and Synthetic CV Syllables', Research on speech Perception, Progress Report No. 10, Indiana University, (1984)

[2] R TUCKER, 'Implementing the Time and Frequency Warping (TFW) Algorithm`, Internal Memorandum, Aston Univerity, (1987)

[3] D L RICHARDS Telecommunication by Speech', Butterworths, London, Chapter 3.4, 173