ASSESSMENT OF TIME FREQUENCY WARPING FOR USE AS A REFERENCE DEGRADATION FOR ASSESSING SYNTHETIC SPEECH

BY

MARTIN DAVID BURRELL

A thesis submitted for the degree of

Doctor of Philosophy

University of Aston in Birmingham

November 1992

ASSESSMENT OF TIME FREQUENCY WARPING FOR USE AS A REFERENCE DEGRADATION FOR ASSESSING SYNTHETIC SPEECH

BY

MARTIN DAVID BURRELL

A thesis submitted for the degree of

Doctor of Philosophy

University of Aston in Birmingham

November 1992

Summary

At present there is no standard assessment method for rating and comparing the quality of synthesized speech. This study assesses the suitability of Time Frequency Warping (TFW) modulation for use as a reference device for assessing synthesized speech. Time Frequency Warping modulation introduces timing errors into natural speech that produce perceptual errors similar to those found in synthetic speech. It is proposed that TFW modulation used in conjunction with a listening effort test would provide a standard assessment method for rating the quality of synthesized speech. This study identifies the most suitable TFW modulation variable parameter to be used for assessing synthetic speech and assess the results of several assessment tests that rate examples of synthesized speech in terms of the TFW variable parameter and listening effort. The study also attempts to identify the attributes of speech that differentiate synthetic, TFW modulated and natural speech.

Keywords:- Time Frequency Warping Modulation, Standard Degradation, Attributes of Synthesized Speech, Perceptual Space, Perceptual Processing

Acknowledgements

I would like to take this opportunity to thank my supervisor Dr. D. L. Richards both for guiding the direction of this research, and for the criticism and encouragement he has given. I would also like to thank Mr. R. W. Whorwood for organizing the financial aspects of the project and his help in the early stages of the project. Thanks to Dr. R. Brewster for his support and fighting my case.

I am grateful to British Telecom for funding the project and special thanks to Mr. B. Surtees for his help and guidance. Thanks also to all the people at Aston university who took part in the tests.

Thanks is also due to my parents, for the support and encouragement they have given me. I am grateful to my friends Linda, John and Jazz for their support and encouragement. I would like to thank Mr. Kaufman and his teams for their medical support plus extra special thanks to the numerous nurses.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Background to Speech Production and Perception

1 Introduction

1.0 General

Telephone information services are increasingly computer based. At present such services require a human operator to act as an interface between the customer and the computer. It is proposed to make such services completely automatic, by enabling callers to communicate directly with the computer via speech recognition and synthesis systems. This study is concerned with the speech synthesis part of such a system. There are an increasing number of commercial speech synthesis systems becoming available for use as output devices. The speech quality of such synthesis systems must be sufficiently redundant to remain intelligible after any degradation introduced by the telephone network. To date the quality of synthesized speech is quite poor. Therefore it is necessary for telephone engineers to be able to rate the quality of speech produced by a particular synthesis system. To determine its suitability for use over a telephone connection.

In the past telephone engineers have used reference devices for rating the performance of a telephone connections. Generally the choice of reference device is governed by the predominant degradation and the type of perceptual errors it produces. None of the reference devices previously used seem particularly suitable for use with synthesized speech. British Telecom Laboratories, Martlesham Heath have proposed a new reference device, based on Time Frequency Warping (TFW) for assessing synthesized speech. This study assesses the suitability of TFW modulation for use as reference device.

1.0.1 The Human and Computer Interface

Modern computers affect all our lives in many varied ways. They prepare our utility bills, bank statements, and store information about us. These systems are mainly data storage and retrieval systems. Other "specialist " systems are being developed that can make medical diagnoses, help plan business strategies and attempt to predict the weather. With all these systems affecting our lives, most people still have very little direct contact with computers. Most computers still require an intermediary between the ultimate user and the computer. Someone who knows how to communicate with the computer via a specialist interface, i.e. keyboard, touchpad, etc. For example if a person wants information on train times, they have to speak, directly or by telephone, to a keyboard operator, who interacts with a computer to retrieve the information. It would be easier if the person requiring the information could interact with the computer directly. In other words speak to the computer. For such a conversation to take place, the computer must be able understand speech input and produce speech as output. In recent years a lot of research has been carried out to make such a conversation possible. This research has been concerned with computer speech recognition processes and computer speech synthesis. For maximum usage most interactive computer information services will be accessible via a telephone link. This study is concerned with the assessment of the quality of the speech output by such systems. The research was funded by British Telecom and is part of their on going research and development of computer information services. Part of the work being carried out at British Telecom laboratories Martlesham Heath, is the development of a reference device to be used in the assessment of text-to-speech synthesis systems. A reference device degrades high quality speech in a controlled and specified way. Experimenters are then able to quantify the performance of text-to-speech synthesizers by directly comparing the test and reference degraded speech to obtain a rating in terms of the reference device. Generally it is desirable for the speech produced by a reference device to be similar to the speech it is being used to assess. The choice of reference device is governed by the form of the predominant degradation and the types of perceptual errors incurred. Hence in the past an adjustable attenuator was used for assessing transmission loss, added speech interference for sidetone, and speech modulated noise was used for quantization effects. This approach enables the subjects taking part in subjective tests to use the same criterion for assessing the reference degraded speech and the speech being assessed. Although comparisons can be made between systems which sound very different, the listener`s task is made much easier if they do sound similar. The use of certain forms of reference device can reduce the variability of subjective scores to a certain extent. None of the reference devices previously used seem particularly suitable for use with synthesized speech now being introduced into public telephone networks. British Telecom have proposed a new reference degradation, Time Frequency Warping (TFW), for assessing synthesized speech. It cannot be claimed that TFW modulated speech sounds exactly like any particular version of synthesized speech but it can be said the resulting displacements in timing do have some resemblance to errors produced by synthesized speech. This study assesses the suitability of TFW. modulation for use as a reference device for assessing synthesized speech. The assessment of synthetic speech is essentially a multidiscipline subject. Therefore before discussing methods for assessing speech communication systems in general and their application to the assessment of synthesized speech it is useful to refer to the mechanisms and processes involved in the production and perception of natural speech.

1.1 Speech

1.1.0 General

The ability to talk is taken for granted by most people, it is something we learn as a child and do not give much thought to the complex processes involved. When we are involved in a conversation we adopt the roles of both talker and listener. As a talker we formulate ideas and concepts, which are translated into a language based code in the mind. This in turn is transformed into an acoustic signal by the vocal organs. As a listener our ears transform the acoustic signal back into a language based code. For computers to talk, the internal representation of data has to be transformed into an acoustic speech signal. This is done via a speech synthesizer, which could be considered as the vocal organs of a computer. A computer's internal representation of a language based code is text stored in digital memory. The synthesizer transforms the stored text into a acoustic signal. All of the transformation processes mentioned above are discussed in more detail so that the complex processes of producing natural and synthetic speech can be compared.

1.1.1 Evolution of Vocal Organs

The human speech organs, originally evolved for purposes such as breathing eating etc. The ability to produce speech is a secondary attribute of these organs made possible by the ability of the human brain to accurately control and manipulate these organs.The vocal organs or articulators for producing speech are shown in figure (1.1).

Vocal Organs

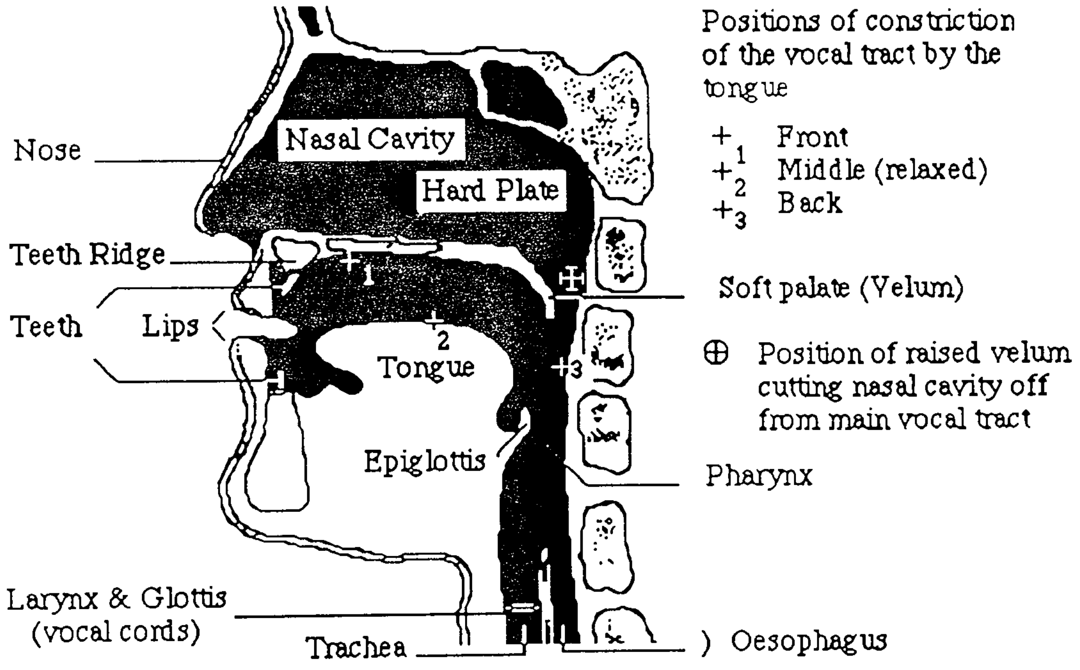

Positions of constriction of the vocal tract by the tongue
+1 Front
+2 Middle (relaxed)
+3 Back
Nose
Nasal Cavity
Hard Plate
Teeth Ridge
Teeth
Lips
Soft palate (Velum)
⊕ Position of raised velum cutting nasal cavity off from main vocal tract
+1
+2
Tongue
+3
Epiglottis
Pharynx
Larynx & Glottis (vocal cords)
Trachea
Oesophagus

Figure (1.1)

Essentially the vocal tract consists of a tube whose resonant properties can be modified using the articulators.(lip, tongue, teeth / jaw). For some sounds the velum is lowered coupling the nasal cavities to the main vocal tract at about it's centre. When speaking the passage from the pharynx to the oesophagus is closed and air from the lungs passes through the larynx (vocal cords). The larynx opens to form a narrow slit, the glottis, during speech the larynx is stretched, and air from the lungs is forced through the glottis which causes the larynx to vibrate opening and closing the glottis. This causes air to be admitted to the vocal tract in a series of rapid pulses averaging about 130 Hz for men and 250 Hz for women, individuals vary by about +/- half an octave. The main vocal tract extends from the larynx (vocal chords) to the lips. For non-nasal sounds (velum up) with the muscles controlling the articulators relaxed and with the mouth slightly open the vocal tract consists of a tube, in this relaxed position the tube is practically constant in cross-section. The length of the tube is approximately 170 mm for men and 135 mm for women. The fundamental frequency F1 of the tube is approximately 500Hz with the second and third resonances at 1.5 Hz, 2.5 Hz etc. These frequencies are known as the formant frequencies and are denoted by F1, F2, F3, etc. The bandwidth of the formants are about 50 Hz for F1 and F2 and about 100 Hz for F3. They vary according to changes in cross-section of the vocal tract caused by movements of the articulators, tongue, lips and teeth. The frequency of vibration of the vocal chords F0 is almost entirely determined by the mass-tension system of the vocal cords themselves, and is relatively independent of the resonant frequencies of the vocal tract. The significance of this is that the coupling between the vocal chords and the vocal tract is so slight as to enable the excitation to be considered separately from the formant system. From the above it can be appreciated that the acoustic signal produced by the vocal organs is complex, containing several frequencies that vary within several bandwidths dependent on the shape of the vocal tract. The relative values of the formant frequencies are dependent on the sound being produced, the absolute values are dependent on the speaker.

1.1.2 Voiced and Unvoiced sounds

Sounds produced by the modulation of the airflow through the vocal chords and the vocal tract are described as "voiced", "unvoiced" or "mixed", which is a mixture of voiced and unvoiced. Voiced sounds are produced by the periodic vibration of the vocal chords when the vocal chords; are open, unvoiced are produced by the turbulent flow of air at a constriction in the vocal tract. The glottal vibrations of voiced sound are considerably modified by the resonances of the vocal tract before they emerge from the lips. The resonant properties of the vocal tract depend on the dimensions and detailed shape of the tract at any particular time, which is determined by the position of the articulators.

1.1.3 Speech Sounds and Language.

The vocal tract is capable of producing thousands of different sounds, but relatively few sounds are used for speech. As mentioned above the shape of the vocal tract and hence the sounds produced are controlled by the articulators. It is therefore convenient to describe speech sounds in terms of the articulatory positions and movements, rather than the sounds themselves. This is the method adopted by phoneticians. The positions of the articulators can be static or dynamic depending on the sound being produced. The fundamental sound units of a language are called phonemes, the number of phonemes used for any given language varies between 20 and 60. Not all languages use the same set of phonemes, which results in some languages using phonemes which do not occur in other languages. For example rolled r's are pronounced in French and by people with Scottish accents, but not in R.P. (received pronunciation) English. A more extreme example is the "tut-tut" sound which is used in some African languages, but not at all in other languages. Another example is in English, we distinguish between / s /and / z /, as in `sip` and `zip` because the change from / s / to / z / changes the meaning of the words. Some languages do not make a distinction between / s / and / z /, but include other distinctions which are not present in English. The English language uses 47 phonemes, which can be reduced to about 40 depending on the dialect, a list of the R.P. English phonemes are listed in app. (A.1). A given phoneme cannot be specified in terms of a unique and unvarying set of acoustic properties (frequencies and amplitudes). Hence the distinction of phonemes in an acoustic signal is relative and not absolute. The acoustic properties of a phoneme can vary quite considerably from speaker to speaker, but the same phoneme will be perceived. The word `boy` can easily be recognized, when pronounced by people of different ages, sexes, and in different voices and accents, although the spectrogram of the phonemes may be very dissimilar. A spectrogram is a visual representation of a acoustic signal.

The acoustic signal which represents a message in a particular language is a concatenation of the characteristic phonetic units. There is some modification of the phonemes when concatenated, the beginning and end of a phoneme are modified by interaction with the phonemes surrounding it. Figure (1.2) shows a simplified spectograph pattern of the phoneme / d / when it is combined with different vowels (di and du). The bars represent the first and second formants.

Spectrographic patterns sufficient for the synthesis of / d / before / i / and / u / .

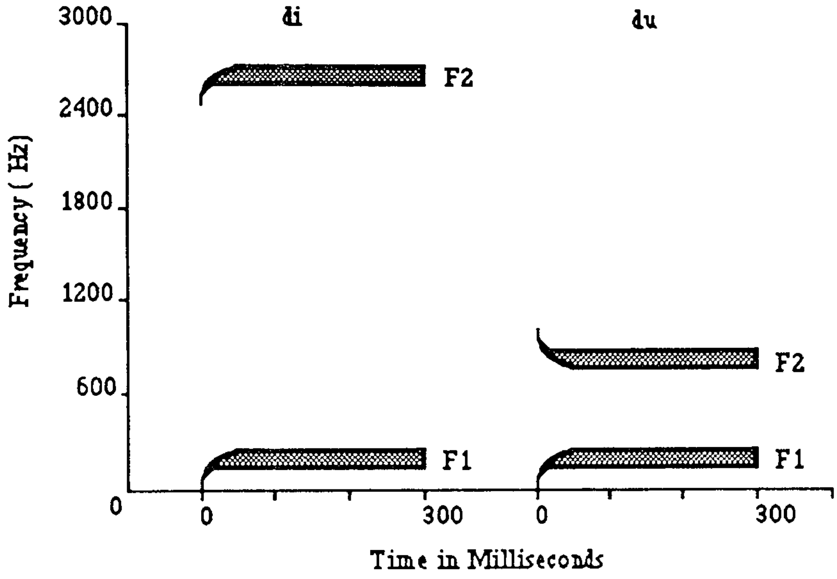


Figure (1.2)

Figure (1.2) shows the second formant at different frequencies and the two types of formant transition. A formant transition is a fairly rapid shift in the position of the formant on the frequency scale close to it`s onset, appearing as a upswing or down swing on the spectrogram. Normally the acoustic cue which carries the most information is the second formant transition. The transition of the first formant (F1), is the same for / b /, / d /, and / g /, and it is the second formant (F2) which distinguishes between them. But, the second formant is quite different for / d / in di and / d / in du, although in both cases the phoneme / d / is readily identified. This lack of an invariable relationship between the acoustic signal and the perceived phoneme characterizes both consonants and vowels. Vowels vary predominantly with the voice characteristics of the speaker, whereas consonants vary markedly as a function of surrounding phonemes, the variations are similar for different speakers.

Phonemes are the smallest acoustic unit that change meaning. The combination of phonemes to form words is specified by the language being spoken. For example a word like "cat" is composed of three phonemes, / c /, / a /, / t /. Changing one of phonemes would produce a different word, like cap, pat, or cut. It should be noted that a phoneme does not necessarily correspond to a letter. "House" is composed of five letters, but only three phonemes, / h /, / ou/, / se / . It should also be noted phonemes do not correspond to a syllable either. "Garden" is composed of two syllables made up of five phonemes, / g /, / ar /, / d /, / e /, / n /.

1.1.4 Vowels and Consonants

The most powerful sounds in any language are the vowels. Vowels provide the structure on which a syllable is based. The vowel provides the core of a syllable, to which consonants are added to modify the sound of the syllable. For example, the vowel / a / is the core of the monosyllable words cat, bat, and nat. The vowel gives the syllable most of it`s acoustic power by providing a relatively long voiced sound, compared to consonants. These voiced sounds are produced when the articulators are static. The differences between vowels are related to the position of constrictions caused by different positions of the tongue (varying the formant frequencies). It has been shown that two degrees of freedom of tongue movement can account for most of the variation in vowels. This simplification implies that, for vowels the tongue is used only in its forward/backward and raised/lowered modes of movement, with the tongue shape remaining fixed. With these two degrees of freedom the vowels can be be plotted in a plane as shown in figure(1.3), it should be noted the jaw also moves.

Vowel Plane with Front and Back Raising / Lowering of the Tongue Axes.

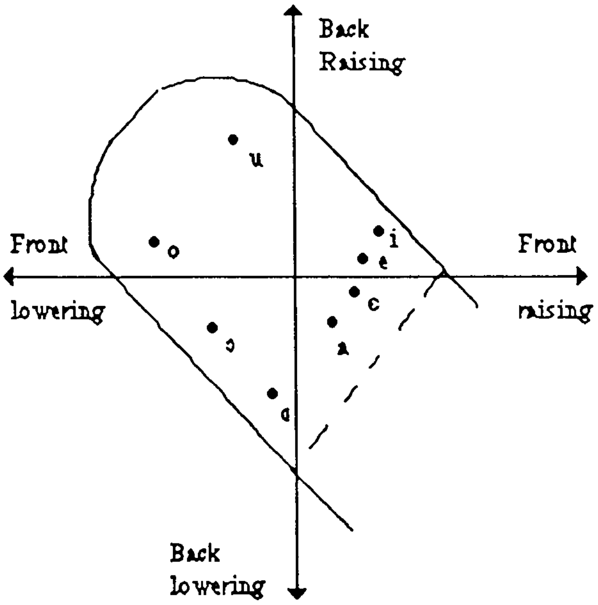


Figure (1.3)

Figure (1.3) is based on physical measurements. The phoneticians' version of this diagram is the "vowel quadrilateral", shown below.

Phonetician`s Vowel Quadrilateral

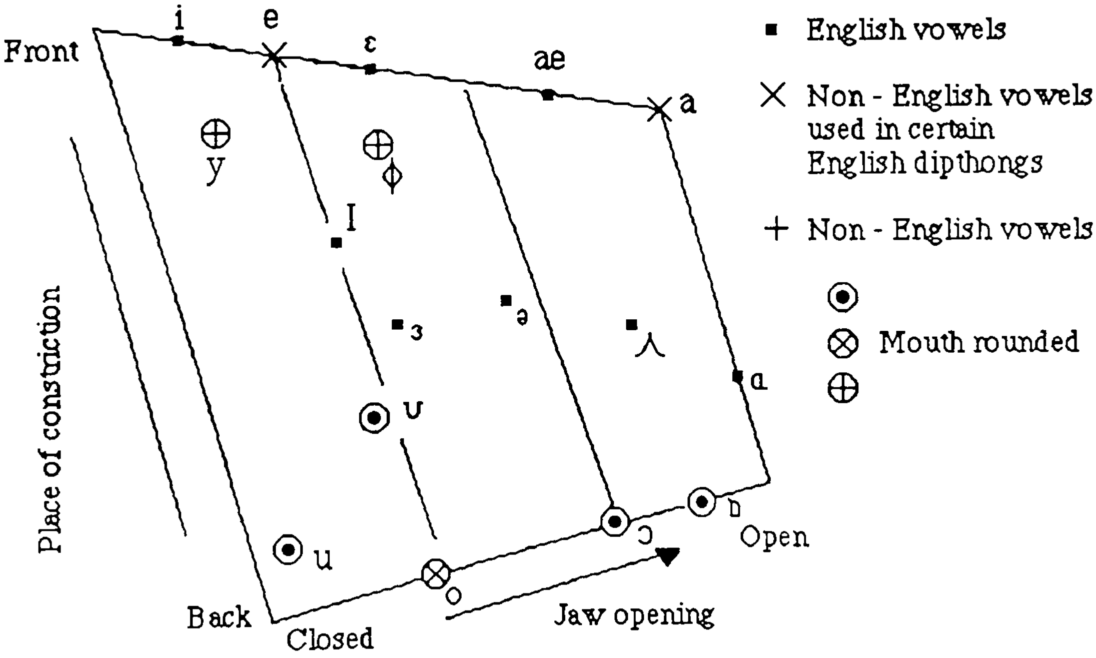


Figure (1.4)

Frequencies of First and Second Formants

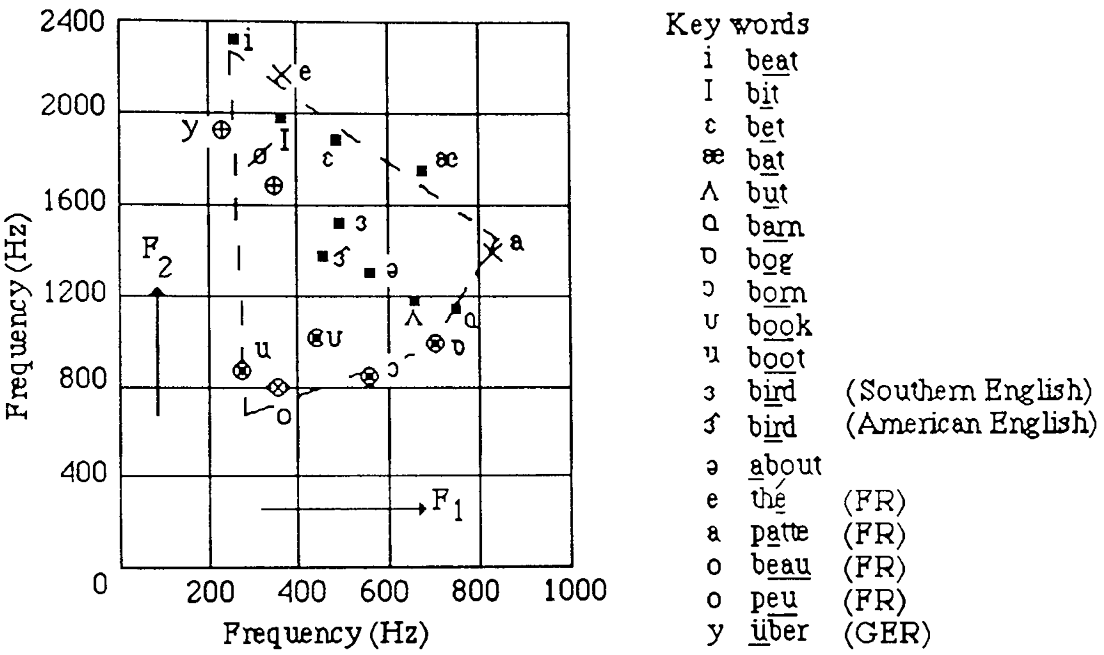


Figure (1.5)

The axes of the figure (1.4) are back / front and open / closed referring to the position of the constriction of the vocal tract and the position of the jaw. The vowel quadrilateral is based on subjective assessment and is a useful representation of vowel quality. Figure (1.5) shows the relationship of the vowels to the first and second formant frequencies, by comparing this to the phonetician`s vowel quadrilateral the relationship between articulatory organs and formants can be seen. There are two additional factors which complicate the characterization of vowels. These are lip position and degree of nasality. These in effect constitute two extra dimensions. Fortunately these two variables do not seem to be used as continuous variables to any great extent. Thus vowels tend to be perceived as either nasalized or not. Similarly only three lip positions seem to have perceptual significance, Pushed out (rounded), normal (lax), drawn back (tense). Thus for each point in figures (1.3) & (1.4) there are six possible conditions of the lip / nasality variables. For example English speakers use the vowels given by the points in figure (1.4) without nasality, but with lip positions varying from tense / i / , to rounded / u /. In contrast , French vowels not only occupy different points on the vowel diagram, indicating different positions of the articulators, but may or may not be nasalized.

1.1.5 Diphthongs Triphthongs and Semi vowels

Closely related to the static vowels are the diphthongs and triphthongs which can be thought as dynamic vowels. These articulatory movements correspond to slow changes between two (or three) static vowels. They can be represented on the vowel diagram as trajectories between static vowel positions. They are used in exactly the same way as static vowels, with the tongue shape fixed and the actual movements slow and smooth. More rapid transitions between static, vowel-like tongue positions give another class of articulatory movements known as "glides", or semi-vowels. In English there are only two, / w / and / j /. The / w / consists of a rapid transition from a / u / position to a / e / position, whereas / j / consists of a rapid transition from / i / to /e/. The glides are not used as vowels (probably because of the rapidity of the transition) and are more accurately thought of as transient consonants. The final class of vowel-like articulations in English are / r / and / l / which are known as "liquids". These two articulations are essentially static but use non-vowel tongue shapes to effect partial closure of the vocal tract. For / r / the tongue tip is curled back to form a narrowing of the tract. In / l / the tongue tip actually makes contact with the roof of the month, allowing the air to pass either side. Both of these articulations vary considerably in detail, depending on language, dialect and content. These two articulations are more properly classed as consonants.

1.1.6 Unvoiced Sounds

So far the speech sounds described are normally uttered using voiced excitation. It is possible, however, to whisper them. For whispered sounds the larynx is held in a loosely constricted position, thereby causing turbulence in the airflow at the glottis. In some languages and dialects whispered versions of the semivowels / w /, / j /, / r / and / l / are valid sounds in their own right. Vowels, semivowels and liquids are produced when the vocal tract is constricted at a particular point, however the constriction does not cause turbulence in the airflow. If the vocal tract is constricted at a point and turbulence occurs, the resulting sounds are known as "fricatives". These consonants exist as voiced / unvoiced pairs and except for / h / are represented by different symbols in the phonetic code. / h / is unvoiced except when preceded by and followed by a vowel as in "ahead".

In this section (1.1) a basic description of the method of how the acoustic signal, (speech) is produced has been given. It should be noted that the acoustic signal is only one of the representations the intended message has gone through. The message originated in the talker`s mind in some form, it was then recoded into neural articulation motor signals which caused the articulators to move in a specific way producing the acoustic signal. The next section (1.2) is a description of the ear which converts the acoustic signal into neural pulses which the listener`s brain decodes (perceives).

1.2 The Ear

1.2.0 General

The ear can be considered to be made up of three parts. The outer, middle and inner ear, are shown below, fig (1.6).

The Ear

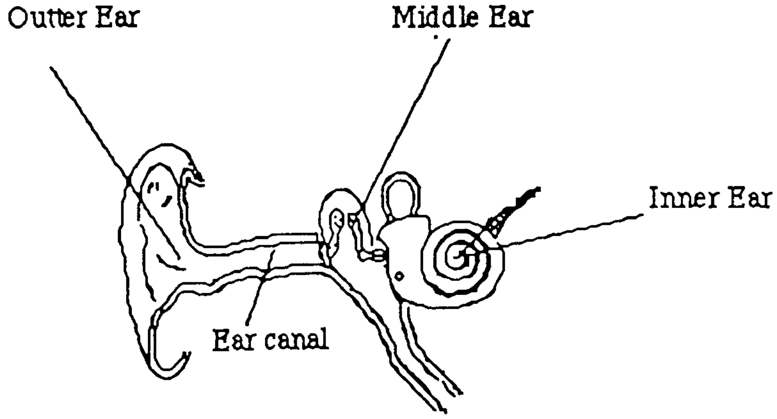


Figure (1.6)

1.2.1 Outer Ear

The outer ear consists of the externally visible ear and the ear canal. The external part of the ear channels the acoustic waves into the ear canal, thus increasing the amount of acoustic energy entering the ear canal. The ear canal is an air filled cavity about 1 inch long, closed at one end by the ear drum, and open to the outside world at the other. The ear canal acts as a acoustic resonator which reinforces frequencies near its resonant frequency. Thus the acoustic pressure at the ear drum for tones near the resonance 3 - 4 KHz may be 2 to 4 times greater than the acoustic pressure at the entrance of the ear canal. This increases the sensitivity of the ear to tones of this frequency.

1.2.2 Middle Ear

The middle ear consists of three small bones which act as levers situated in a air filled cavity, figure (1.7). The cavity is connected to the back of the mouth by the Eustacian tube. The Eustacian tube protects the ear drum from damage by equalizing any difference in air pressure between the ear canal and the middle ear cavity.

Middle Ear

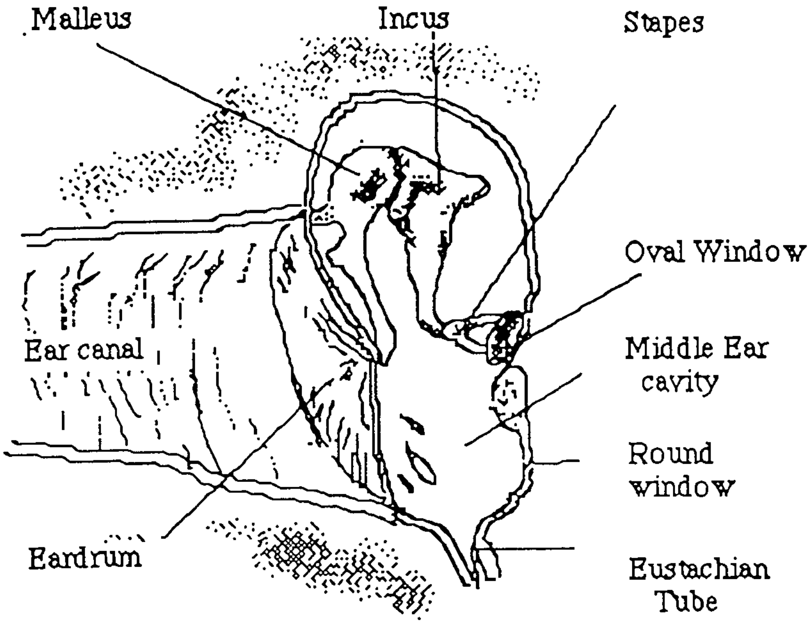
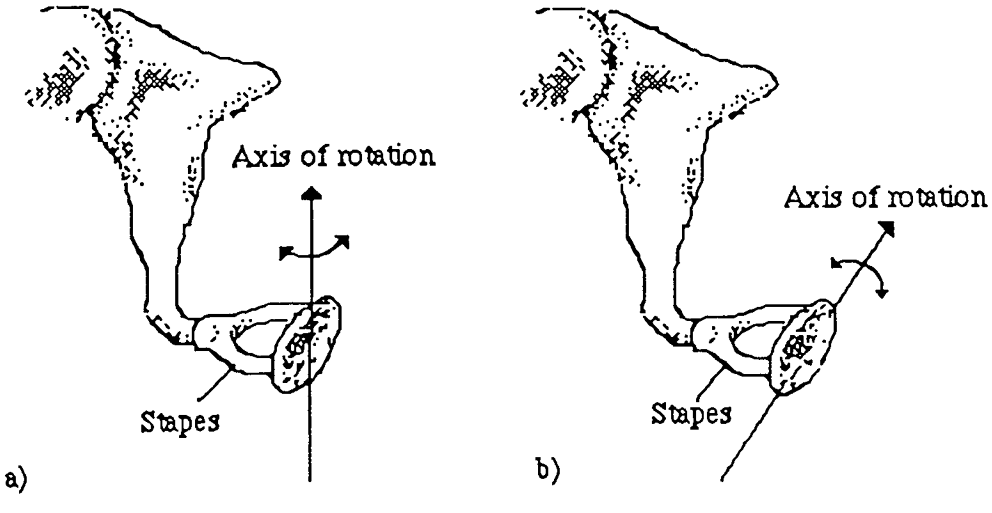


Figure (1.7)

The three bones form a system of levers which couple the 8.5 x 8.5 mm eardrum to the 3.2 x 3.2 mm membrane of the oval window, which is part of the fluid filled cochlea or inner ear. The three bone coupling magnifies the force at the eardrum 30-60 times, with a corresponding reduction in displacement. Under normal listening conditions the motion of the eardrum is very small, however for loud noises the motion is greater. If the middle ear was a "fixed" coupling between the eardrum and oval window, the oval window could be damaged by loud noises. To protect the inner ear from the effect of loud noises, the efficiency of the middle ear is reduced. This is achieved by the coupling of the malleus and incus dislocating when responding to loud noises. Large amplitudes of the eardrum also causes some non-linearity in the force at the oval window of the cochlea. The change in coupling for loud noises is shown in figure (1.8)

Normal and high intensity modes of vibration of the stapes.



a) normal mode b) high intensity mode Figure (1.8)

1.2.3 Inner Ear

The inner ear or cochlea is made up of a helical tube which is divided along its length. At the "head" of the cochlea there are three semicircular canals set mutually at right-angles. The fluid in the three canals is connected to the fluid that fills the cochlea. The three semicircular canals are connected to nerve endings which are stimulated when there is any change in the position of the head. The three canals are associated with balance and are independent of the hearing mechanism. The two chambers of the cochlea are divided partly by a bony ridge and partly by the basilar membrane and are connected by a small hole called the helicotrema. The helicotrema is situated at the apex of the spiral furthest from the oval window. The chamber from the oval window to the helicotrema is called the scala vestibuli, the other chamber, called "scala tympani" connects the helicotrema to the round window. Which is a flexible membrane separating the fluid in the scala tympani from the air in the middle ear cavity. The basilar membrane supports the organ of Corti.(organ of hair) . Which consists of four rows of hair cells, the nerve endings of the cells are connected by thousands of nerves to the brain. The Corti organ acts as a spectrum analyzer by providing frequency and intensity information. The vibrations of the oval window are propagated through the fluid in the cochlea. This causes transverse displacement of the basilar membrane. The maximum displacement of the basilar membrane occurs at different places along it's length according to the frequency of the oval window vibrations. The velocity at which the wave is propagated along the basilar membrane is such that the wave takes 5ms to reach the helicotrema which is 30-35mm from the oval window. The relationship between the point of maximum displacement of the basilar membrane and the frequency of vibration of the oval window is shown in fig (1.9).

Relationship Between Maximum Displacement and Frequency of Vibration of the Oval Window

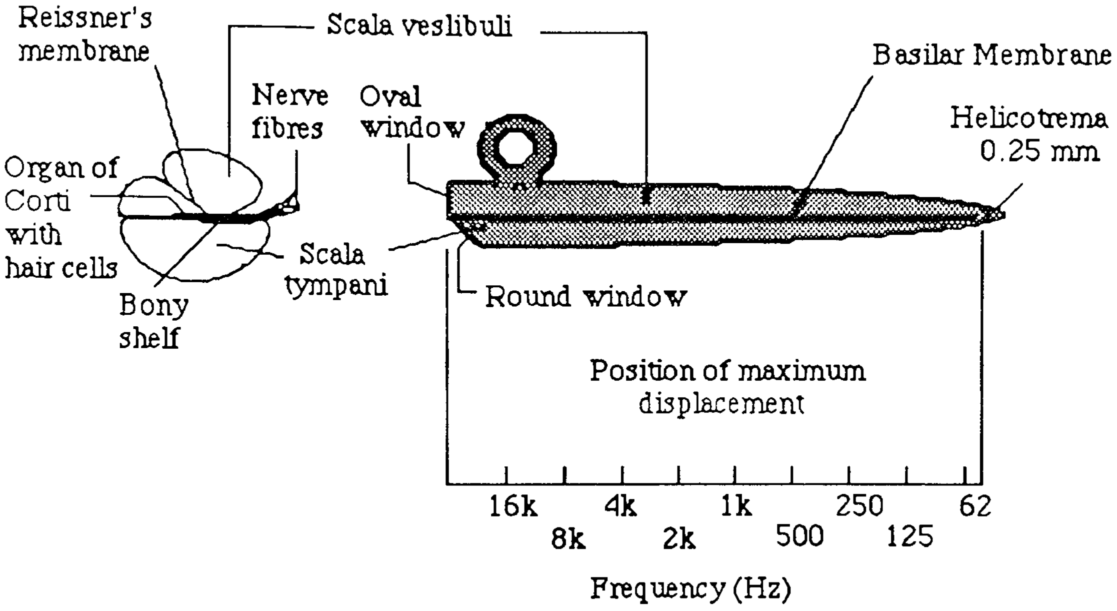


Figure (1.9)

It can be seen that the higher frequencies stimulate the hair cells near the oval window and the lower frequencies are detected nearer the helictrema. In this way the nerve fibres are able to transmit frequency information about the oval window vibrations, without having to signal to the brain at a rate higher than 100Hz, though spike rates can be higher. This is because the hair cells are sending the position of displacement not the number of vibrations of the oval window per second. The analysis performed by the cochlea applies for sound frequencies down to just below 100 Hz. Lower frequency sounds than this are just audible, but it is believed that the sounds modulate the periodicity of the nerves impulses directly.

The cochlea also has to send to the brain information on the intensity of the nerve stimulation. The intensity of the stimulation is indicated partly by the number of pulses per second and partly by the range of sensitivities of the hair cells at the point of stimulation. This arrangement means there is a threshold of hearing. The minimum sound intensity that can be detected, has to be able to just stimulate the most sensitive hair cell at the particular point on the basilar membrane. Where the physical detection of sound is concerned the ears act independently. There is a 50dB attenuation through the skull bone. Thus any sound that comes from the right is detected by the right ear first and the left ear a short time later. The correlation between the two stimuli give the listener an indication of the sound direction.

1.3 Cognitive Processing

1.3.0 General

Cognitive processing is the name given to the process that takes place in a listener`s mind, which interprets the neural information from the ear. The ear converts all acoustic input into neural information . The neural information is fed to cognitive / short term memory which filters out useful parts and decides whether action or further processing is necessary. Under normal circumstances this process is carried out automatically at a subconscious level. But if someone calls your name, say in a noisy room this is immediately brought to the attention of the conscious mind. The cognitive process took a sequence from the temporal neural signal and matched it with an internal representation of the listener`s name. The listener is then said to have perceived his or her name. Exactly how the acoustic / neural signal was processed and matched is not fully understood. There have been various theories put forward by researchers working in the field of cognitive psychology to unravel the very complex process of speech perception. The size and type of the internal representation of speech can affect the perception of speech and the "meaning" perceived. This implies the the use of a perceptual unit.

1.3.1 Perceptual unit in Speech Perception.

In speech perception the problem of segmentation arises. Confronted with a continuous flow of speech, the listener must make decisions over segments of the signal as it comes in to determine whether or not the segment is a perceptual unit of speech. But what constitutes a perceptual unit ? Under normal listening conditions longer pauses can occur in speech within a word than between words. Therefore it is not a simple matter of determining pauses between words.(under adverse conditions the talker may supply word or even letter segmentation). Hence for a listener, listening under good conditions must impose his own segmentation based on his own perceptual unit.The size of the perceptual segment is dependent on the capacity or efficiency of the processing mechanism (cognitive memory), and the listening conditions. Since many of the acoustic cues for speech perception involve the analysis of temporal properties, the information must be accumulated and held long enough for the temporal relations to be integrated. This is achieved by temporary holding the auditory information in a buffer store called echonic memory. The duration of the echonic memory has not yet been accurately determined. If we accept one estimate that it persists for 2 seconds, then at a speech rate of 120 words per minute, it`s capacity would be about 4 words. The signal must be decoded and transferred to permanent store before the echonic trace has faded. If the decoding of the signal is delayed for more than 4 words, that part of the signal would be lost. It follows that if this estimate of echonic memory duration is correct , the maximum size of the speech perception unit must be about 4 words which is close to the size of a typical phrase. However, it is possible that the echonic memory can hold information for longer than 2 seconds. Other research has suggested a duration up to 9 seconds Ref(13) so the maximum size of the speech perception unit could be larger. In section (1.1.2) it was stated that the phoneme was the smallest sound unit that could change meaning. Hence it could be argued that the phoneme is the unit of perception. Below there is a brief description of some experiments which put forward some possibilities for a unit of speech perception.

1.3.2 Identification of Speech Perceptual units

Speech interruption experiments based on the assumption, that the maximum disruption in the perceived speech will occur when the perceptual unit is split were carried out by Huggins ref (9)1964. Huggins found that interruptions that split the syllable were most detrimental, and hence concluded that the unit of perception of speech was the syllable. Steinhieiser and Burrows (1973) ref (33), also proposed the the syllable as the unit of perception. They compared reaction times for making judgements about consonant - vowel - consonant trigrams in a variety of tasks;

1 detection of final consonant

2 determine if the trigram is a real word

3 judge two trigrams to be physically identical

4 judge if two trigrams are real words

1 and 2 took about the same amount of time, indicating that phoneme identification does not precede word identification. But 3 was faster than 4, which indicates trigram identification precedes word indentification. In another set of experiments subjects listened to sentences, and the time to identify target items was recorded. Subjects had to press a key when they detected the target item.The results showed that the detection of phonemes was slower than the detection of syllables or words. For example the word `snow` was be recognized faster than the phoneme / n /. This is contrary to what would be expected if speech was perceived phoneme-by-phoneme. Warren and Obusek ref.(38) carried out further experiments, where a single phoneme in a sentence was replaced with a noise (a cough). It was found that subjects did not perceive the change. This phenomenon "phoneme restoration" provides further evidence that speech is not normally perceived phoneme-by-phoneme.

1.3.3 Shadowing Experiment

A shadowing experiment ref.(13) (Liberman et al) showed that although phonemes could be discriminated, speech is not normally processed phoneme-by-phoneme. In shadowing experiments subjects reproduce speech they are listening to, following the input as close as possible. In some rare cases skilled shadowers were able to reproduce speech only one phoneme behind the input speech. This was taken to mean that these subjects were processing the input speech phoneme-by-phoneme. In such cases the subjects had no comprehension of the message. Obviously the subject`s performance was not analogous to normal speech perception.

1.3.4 Bottom-up Top-down Perception of Speech

For the meaning of a speech utterance to be understood a listener must draw on several knowledge bases. These knowledge bases which have been built up over the listener`s life time are stored in the listeners memory. They can be classified broadly as language knowledge and world knowledge.

Language knowledge includes:-

Sounds (phonemes) and spellings of words
Semantics - normal meaning of words
Syntax - grammatical rules for acceptable utterances
Idiomatic and metaphorical interpretation

World knowledge includes:

How things operate and relate in the real world
Specific knowledge of people or events

In the perception of speech some or all of these knowledge bases are used to perceive the "meaning" of the utterance. There has been a lot of debate about the perception process centred on whether it is a bottom-up, top-down process or a combination of the two. If we take for example the spoken sentence "He stopped when he saw the red light.", which is part of a longer discourse i.e. reading aloud from a book. A bottom-up process would process each phoneme individually to form each syllable / word, check that each word and sentence was syntactically and semantically correct and that the utterance was in context with the rest of the discourse in a world sense. A top-down process would start with a expected sentence structure based on world knowledge, language knowledge and context information. The most likely process is a combination of the two processes bottom-up and top-down. A simplified illustration of this process is shown in figure (1.10).

Bottom-up Top-down Perception

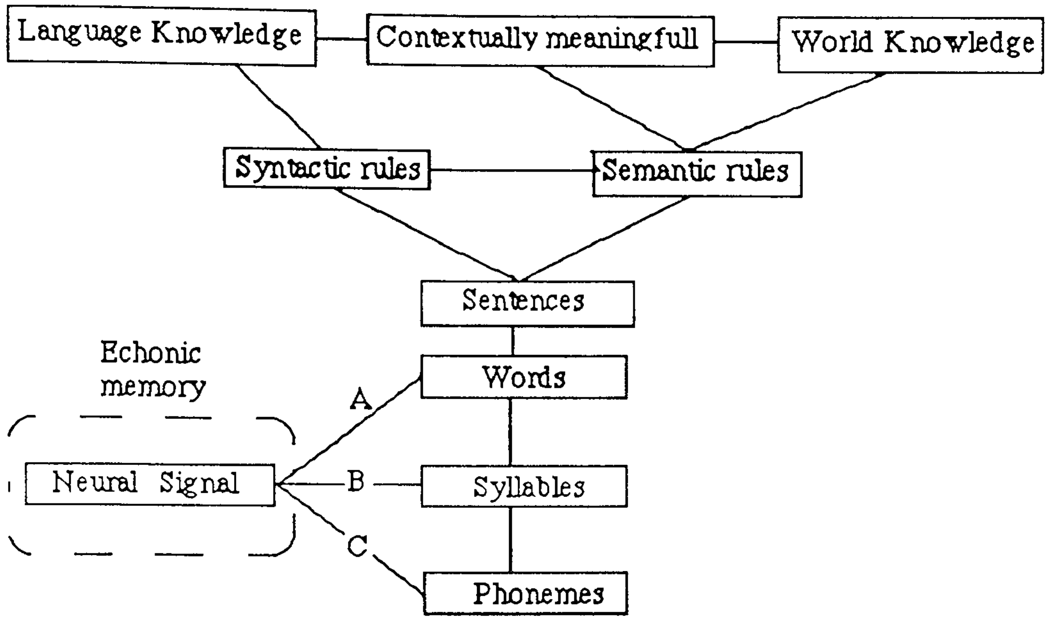


Figure (1.10)

If a listener is listening to speech which is part of a larger discourse, i.e. a speech or an, on going conversation. Then there will be additional information to the acoustic signal which the listener can use to aid the perception of the speech. The additional information will be made up of semantic information (word meanings) based on contextual information from previous parts of the discourse and the listener's own world knowledge. This additional information combined with the listener's language knowledge gives the speech a degree of predictability. The degree of predictability is dependant on the individual listener, the quality of the acoustic signal and listening conditions. The predictability of the speech affects the amount of effort required to understand the speech and may affect the internal processes used in the perception of speech. For example the perception of C-V-C words presented as part of a rhyme test, which requires the listener to identify individual phonemes. The internal perception process will require detailed processing at the phoneme level. There will be no world knowledge, contextual information and minimal language information. In fact the only language information will be determining whether the perceived sound units are used in the language and whether it is a valid word. In contrast the perception of high quality speech based on a familiar topic may only require cursory analysis of the signal stored in the echonic memory. This may take the form of only identifying the general characteristics of the pitch contour. This combined with the "expected" phrase based on a combination of world knowledge, language knowledge, context, syntactic and semantic rules. Maybe sufficient information for the correct perception of the speech. If the general characteristics of the pitch contour differ significantly from the "expected" phrase pitch contour. The "expected" phrase will be modified or abandoned, at the same time a more indepth analysis of the stored signal is taking place. It is possible that the cursory and indepth analysis started at the same time and run in parallel. If we consider figure (1.10) three levels of analysis A, B, and C are carried out in parallel. The time each analysis takes, increases from A to C, if the top down analysis matches the bottom up analysis at any level, the signal will be perceived. The actual perceptual strategies adopted by a listener will be dependant on many factors, which include the quality of the acoustic signal and the listening conditions.

1.4 Development of Synthesized Speech

1.4.0 General

The first truly electronic speech was synthesized by the invention of the vocoder (voice coder).The vocoder analyzed speech into low bandwidth spectral components, and recombined then to reproduce the original speech. The purpose of the vocoder was to reduce the transmission bandwidth of speech, for transmission over a telephone connection. Dudley's vocoder used a ten channel filter bank to provide spectral analysis of the speech and a circuit to detect the pitch of voicing. The speech was resynthesized by using pulses at the pitch frequency to excite a set of ten filters, the gains of the filters were proportional to the values of the ten spectral components. The synthesis half of the machine was later modified to permit manual control, so that it could be "played" via a system of keys and switches. This machine was demonstrated at the 1939 world fair. Vocoder research produced other synthesizers, but manual operation made the accurate specification of control parameters virtually impossible. One of the methods adopted to overcome the problem, was to paint the control parameters onto transparent tape, and used light modulation to convert these to electrical signals. The machine of Lawrence (1953) ref (12), and the pattern play-back devices Haskins laboratory (Cooper, Liberman & Borst 1957) ref (5) are typical examples.

From the 1950's onwards, the availability of computers provided a more systematic way of producing the control parameters of the synthesizers. Computers enabled researchers to model the hardware synthesizers, hence enabling them to optimize the synthesizers under development before they were built. In 1964 J.N.Holmes, J.G. Mattingly and J.N. Shearme produced some "excellent" speech generated by rule using a formant synthesizer. The synthesizer used resonant analogue circuits to model the resonances (formants) of the vocal tract.

By the 1960's it became apparent that there were two different approaches to the design of speech synthesizers. One approach models the speech production mechanism, "articulation synthesis". The other approach models the speech signal. The two approaches to a certain extent reflected the division of interests occurring at the time. Articulation synthesis was and still is of interest to researchers interested in developing a greater understanding of the speech articulation mechanism. The speech signal model approach, was still part of vocoder research, which involved reducing the number of bits required in the encoding, transmission and resynthesis of a speech signal.

Around 1970 speech synthesis for computer output became a more important stimulus for research than vocoder design. Also at this time vocoder research began to fall off, for two reasons. One new technology had made bandwidth in telecommunication systems cheap and plentiful. The second was for most purposes, the vocoder problem had virtually been solved, by the use of Linear Predictive Coding (LPC.). In 1971 B.S.Atal and S.L.Hanser ref (2) published a paper that showed LPC could be used to analyse and resynthesize speech. LPC techniques model the speech wave form, by modelling the actions of the vocal tract with a multi-pole digital filter. The parameter input data to the filter, consists of an energy parameter, a pitch parameter, and a set of filter parameters. The overall bit rate can be as low as 800 bits/sec or as high as 9.6 Kbits/sec, which is dependant on the speech quality required. LPC provided an algorithmic technique for the analysis resynthesis of speech which was well suited to digital implementation. The low number of bits required to store or transmit the speech parameters made LPC the first choice for computer speech output devices.

In the 1970's large scale integrate circuits (LSI) gave researchers relatively cheap and increasingly powerful digital computers. The two factors LPC and LSI stimulated a prolific increase of speech synthesis research. The analogue components of speech synthesizers were either simulated by computer programs or replaced by digital circuitry. A digital version of formant speech synthesis was developed. The input data represented the mid-frequencies and the bandwidths of each of the vocal tract resonances (or Formants) and the pitch frequency. The data rate can be similar to that of LPC.

1.4.1 Rule Synthesis and Analysis-resynthesis

It is useful to distinguish between analysis-resynthesis or (Vocoding) coding, from rule synthesis. Analysis-resynthesis starts with a spoken message which is analysed, for example by a bank of filters (formant coding) or LPC, followed by some form of data reduction, the reduced data is then either stored or transmitted. At the 'receive end' the speech signal is resynthesized. At low bit rates the reduction of speech quality can be considerable but at high bit rates it is negligible. Where as a rule synthesizer generates the output speech from a set of rules based on text. A rule synthesizer can be divided into three main components, text input, parametric dictionary, and speech output device. The first component, depending on the computer system and its application, uses some form of text input which is transformed into sound units. The second component uses the parametric dictionary to look up the parameters of each sound unit. The sound units are concatenated, smoothed, and output via the output synthesizer.

The widely used speech output units are based on either LPC or formant synthesizers. In the application of speech synthesis systems there is a trade off between analysis-resynthesis of fixed texts of high quality speech, and rule synthesis of unlimited text, and generally lower quality speech. In this study the work is confined to rule text-to-speech synthesis.

1.4.2 Three main types of rule synthesizers.

1.4.2.1 Articulatory based systems

Articulatory based synthesis is mainly of research interest. These synthesizers are sophisticated models of the human vocal tract. The control parameters are specified in terms of voice and articulation mechanism, through which natural boundaries and interactions are preserved.

1.4.2.2 Diphone based system

In this study "diphone - based synthesis" refers to systems which use prestored speech fragments, where they can be either diphones, demisyllables or syllables etc. The speech fragments generally are segmented from natural utterances spoken by a trained speaker and stored in parametric form, like formant or LPC parameters. By smoothed concatenation of these fragments it is possible to generate speech of quite high quality, without much processing. The disadvantages of such systems is that they are restricted to one voice, if another voice is required, a complete new diphone vocabulary has to be complied. Further more it is not easy to introduce unstressed syllables, rate changes, or changes in articulation quality. Due to the fact the these features are already established in the speech segments, based on the segments position in the original utterance and the speakers voice. To attempt to change any of the characteristics of the speech can result in unnatural transitions within the output speech.

1.4.2.3 Allophone based system

Allophone synthesis has the greatest potential for producing very high quality synthetic speech. All of the transitions, i.e., text to phonetic code, phonetic code to speech parameters, are controlled by rules. Every phoneme of a language and their allophonic variations are stored in parametric form. The transformation rules control the allophones pitch, duration and changes due to stress, speaking rate etc. Although at present the quality of the speech produced by these systems is fairly poor, further refinement of the rules should produce high quality natural sounding speech.

1.4.2.4 Synthesizer Data Requirements

The motivation behind most research into synthetic speech is to produce systems which will not require vast amounts of memory in their implementation, while maintaining a high quality speech output. At present there is a trade off between speech quality and memory size. Figure (1.11) taken from ref.(4) shows the data rates associated with various speech synthesis methods. The axis also shows how many English words of average length (0.6 seconds) could be stored in 128K bit ROM at each data rate.

Memory Requirements of Different Synthesis Systems

Data Requirements of Synthesisers

Figure (1.11)

1.5 An Example Speech Synthesizer

1.5.0 Overview of MITalk text-to-speech synthesizer.

An overview of the MITalk text-to-speech synthesizer has been included because the work carried out in it`s development has formed the basis of a lot of research into speech synthesis in the the U.S.A.. This overview was adapted from ref (1) also shows the essential transformations the original text message has to go through to produce speech. These transformations are common to all text-to-speech synthesizers. MITalk was developed over a number of years from the early 1970's to the early 1980's by Jonathon Allen and Dennis Klatt . The MITalk system is a allophone based rule text-to-speech synthesizer. It is made up of 9 independent modules which allows the user to update or redevelope a module to suit their own needs. The modules can be divided into 2 groups, modules 1-to-5 are the text analysis modules and modules 6-to-9 are the speech synthesis modules. Figure (1.12) shows the order in which the text input is processed and how groups of modules can be used as speech output devices for input data of different formats.

SCHEMATIC REPRESENTATION OF MITalk

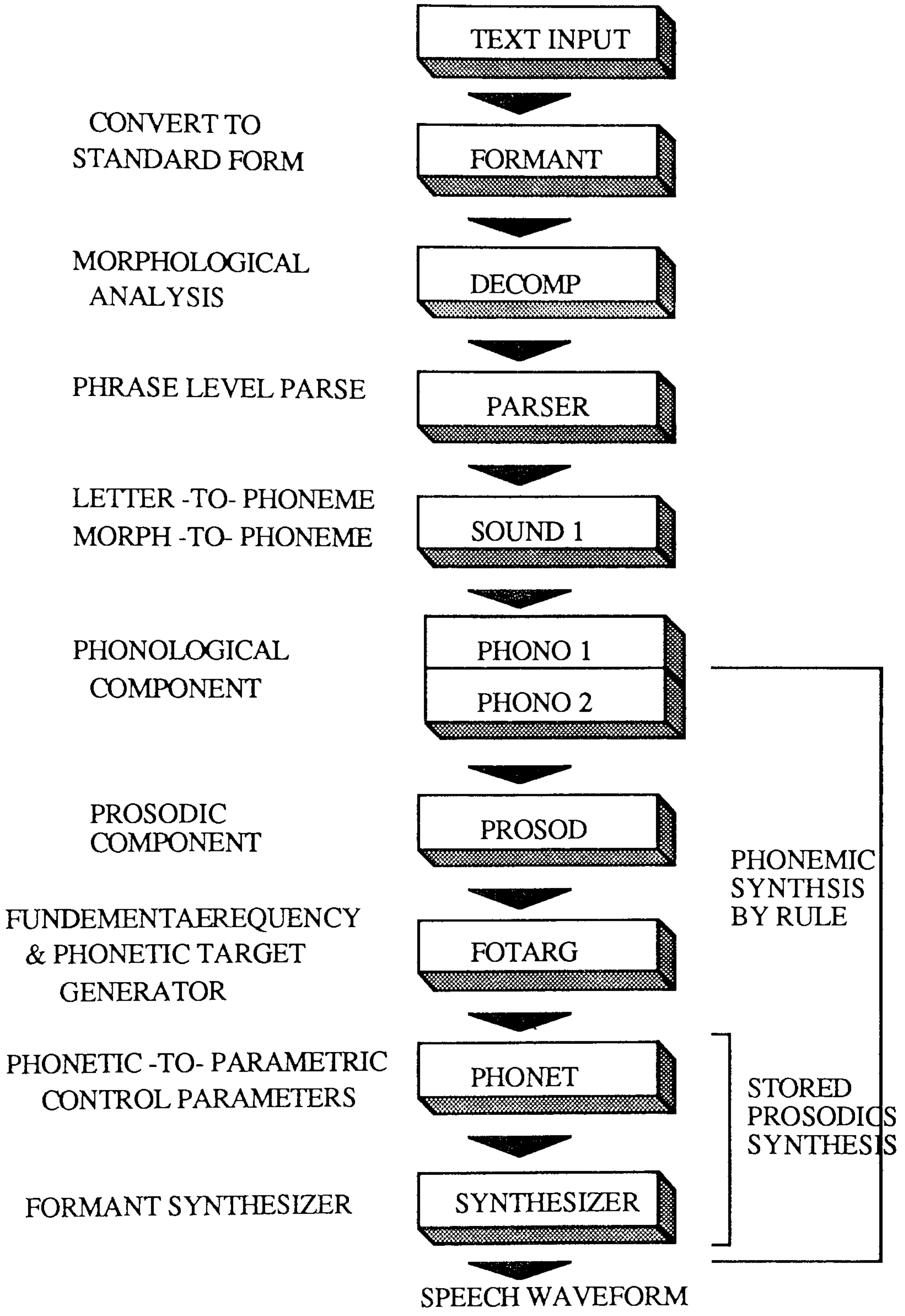


Figure (1.12)

1.5.1 MITalk Analysis of text

Module-1 FORMAT

Unrestricted text is input via a keyboard, symbols within the text are converted to standard form, i.e. Mr -> Mister, 6 -> Six.

Module-2 DECOMP:- Morphological Analysis

The standardized text is broken down into morphs. A morpheme is the minimum meaningful syntactic unit of a language. When a morpheme is represented by a letter string segment, it is called a morph, they can be used as a basis for determining word pronunciation. For example the word "snow" is one morph, "snow-plow" is a compound of two morphs and snowplowed is made up of two root morphs and a suffix. MITalk has a 12,000 morph dictionary which contains the spelling, pronunciation and part-of-speech information for each morph. If a letter string cannot be analyzed then the letters are represented separately. The part-of-speech information indicates the relative peak of the fundamental frequency due to the particular morph, i.e. if the morph is an article, relative pronoun, interrogative adjective etc. The relative peak of the fundamental frequency for an article equals 0, and for a interrogative adjective it would be 4.

Module-3 PARSER:- Phrase Level Parsing

Phrase level parsing (breaking down the sentence) is performed to aid selection of prosodic correlates, by providing a surface structure of the sentence.

Module-4 SOUND 1:- Phonetic Transcription

The morphs and letters are transformed into phonetic code. The output from SOUND 1 consists of segment labels, stress marks, syllable and morph boundaries of a word. Some recoding is done on the phonetic transcription, based on constant "flapping", insertion of glottal stops, and selection of alternate pronunciation of "THe". The sentence phrase information is output to the phonological component.

Module-5 PHONOL:- Phonological Component

PHONOL is divided into two parts, Phono 1 handles the phrase information. The effects of suffixes, and compounding on lexical stress are computed, allowing the use of stress marks in the transcription and changes in vowel colour. Semantic effects due to particular lexical items, such as negatives, are found. These have important effects on pitch. Phono 2 recodes the phonetic code into allophones to account for any interactions between phonemes.

1.5.2 - Synthesis of Speech

Module-6 PROSOD:- Prosodic Features

This module uses the sentence representation output from Phono 2, each speech segment is assigned a stress feature. For each speech segment the basic duration, prepausal lengthening, pause duration, and pollysyllabic shortening, and the effect of clusters are determined.

Module-7 FOTARC:- Fundamental Frequency Contour Generator

This module uses information from the syntax and phonological components. It uses the phrase of each sentence analyzed by the parser to determine the declination line through each phrase and to insert continuation rises. Lexical stress marks and syllable division are used to determine the location of the fundamental frequency peaks, and part-of-speech information is used to determine the relative height of the peaks. Phonemic data provides the information needed to determine segmental influences on the fundamental frequency. The process utilizes a context window five segments wide. There are twenty such parameters that vary with time. Given the prosodic frame work, phonetic targets are determined for each phonetic segment, these are two values of the fundamental frequency, one at the onset of a speech segment and the other is the mid-value

Module-8 PHONET:- Parameters Conversion

This section, accepts input from the fundamental frequency module in the form of an array of phonetic segment names, segmental stress feature, segment duration , and the two fundamental frequency targets for each phone. The output of this section produces 20 synthesizer control parameters every 5 msec.

Module-9 Hardware Formant Synthesizer. (Wave form Generation)

The terminal synthesizer uses the 20 control parameters (updated every 5 msecs) to generate the speech waveform. A special purpose hardware formant synthesizer was used to perform this task in real-time. The digital speech signal is generated at 10 KHz, and then converted to analogue form via a D/A converter and low-pass filter.

A number of versions of the MITalk system have been developed, PROSE 2000 (Bernstein & Pisoni), PROSE 2020, PROSE V3.0, Klattalk (1982) and DECtalk (Bruckett, 1984). Some of these systems have been commercialized. The DECtalk system has 9 voices, one of them Perfect Paul can produce comparatively high quality natural sounding speech. MITalk and its variations produce speech with an American accent, Dennis Klatt had developed the allophones for the MITalk system based on extensive analysis of his own voice, hence the characteristics of his American accent were transferred to the allophones.

Chapter 2

Review of Speech Assessment Methods

2 Assessment of Speech Communication Systems

2.0 General

Speech assessment methods can be divided in to two general groups, absolute and comparison methods. Comparison methods rate the test system by adjusting a standard system until the test subjects perceive the two systems to be equal. The test system is rated by the equivalent setting of the standard system. Absolute methods determine quality according to a particular criterion, i.e. listening effort.

2.1 Speech Paths

2.1.0 General

The results of speech quality or communication efficiency tests made on a speech path are only valid for the particular speech path tested. Hence any results made on another system can not be directly compared unless the second path`s physical specifications are equivalent to that of the first.

2.1.1 Fundamental Speech path

The most fundamental speech path is an air path between two people participating in a face to face conversation. The term "speech link" is used to denote the complex transmission paths involved when a conversation is in progress. When measuring speech quality, it is the efficiency of the complete system that must be measured . For a face to face conversation there are several characteristics that must be considered. Figure (2.1)

Basic Paths Involved in Face to Face Conversation

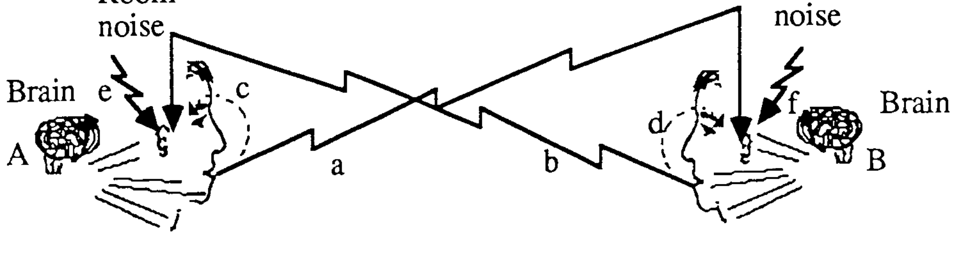


Figure (2.1)

Paths 'a' and 'b' represent the speech emitted from the participants mouths. 'c' and 'd' are the side tone paths, by which participants hear their own voice. Any room noise present will affect all these speech paths. Room noise reaches the participant's ear via the paths 'e' and 'f'. If the room noise masks the speakers own voice paths 'c' and 'd', this will cause the speaker to talk louder. The same effect can be observed when somebody is listening to loud music on headphones, they tend to speak louder because the music is masking their own speech 'c'. If the room noise masks the other speakers voice, this will make the conversation more difficult. When the conversation takes place over a telephone connection the speech paths mentioned above are still relevant, but will have become more complicated due to the telephone apparatus and the lines through which the speech passes. Any room noise will reach the participant via the ear not being used for telephoning, leakage under the earpiece and via the microphone through the telephone's own side tone. Other attributes of the telephone connection such as attenuation, noise, cross talk and spectral shaping will affect the speech quality over a telephone connection.

2.1.2 Definition of 1 metre Fundamental Speech Path

Essentially a reference system is a speech link composed of stable and specifiable components, intended for laboratory conditions where the variables affecting the speech link can be defined in simple physical terms. An important factor in the design of a reference system is that the fundamental principle on which it is based must not restrict the design of the components (Earphones, microphones, etc.). The fundamental principle on which the current internationally used reference system ARAEN, is based on a one metre air path, fig.(2.2)

Fundamental Principle of One Metre Airpath Reference System

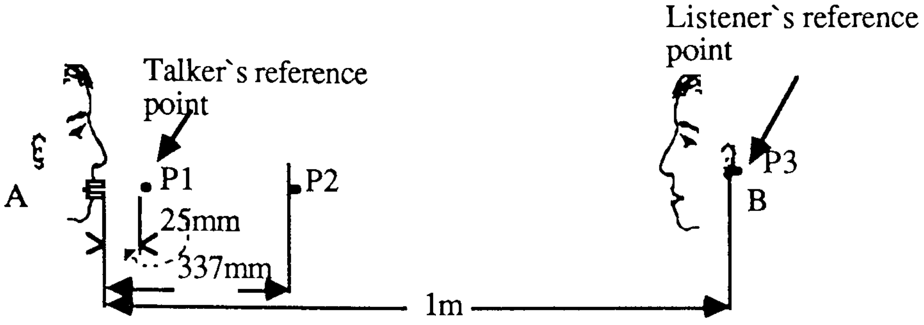


Figure 2.2

2.2 Reference Devices

2.2.0 General

Reference devices developed from the need of telephone network planners to determine the effects of different components on transmission quality. A reference device has an adjustable component which introduces a specific controlled degradation between a specified "send end" and a specified "receive end" of a reference system. The main concern of telephone planners in the early days was transmission loss or lack of sensitivity. In 1925 an adjustable attenuator which connected the "send" and "receive' ends of a reference connection was standardized. The equipment became known as the"European Fundamental Telephone Reference System". Comparison could then be made between an "unknown" telephone connection and the reference connection using ratings called reference equivalents. Reference equivalent measures determined equivalent overall mouth-to-ear sensitivity of the speech. Such a reference device is still in use today although the hardware of the reference system has been up-dated and the method of specification modernized. Other controlled degradations, such as added circuit noise, sidetone random speech interruptions and modulated noise reference devices have been developed over the years. This allowed the telephone network planners to determine trade-offs between, for example a given level of circuit noise or a given amount of attenuation distortion and an equivalent change in overall transmission loss. As already mentioned all the reference devices are made up of a controlled degradation and specified send and receive systems. These systems are usually referred to as reference connections (speech paths). The specification of a reference connection is in terms of stable and specifiable components which may be based on a fundamental speech path.

2.2.1 Development of Reference Systems

The problem of comparing different speech systems which introduce various degrees, or types of degradation into the speech can be overcome by the use of rating measurements. Ratings are defined as equivalent settings of a given reference system. The principle of such tests, is that each speech system is directly or indirectly compared with a reference system, using loudness articulation or any other specified criterion of equivalence. The results yield a single number rating the test system`s speech quality or communication efficiency. A reference system consists of a reference speech path where the parameters of the path are specified used in conjunction with a reference device (controlled degradation). Controlled degradations are simple, readily specifiable and repeatable scales of measurement, against which unknown or more complex phenomena can be rated. The rating numbers may not be in themselves subjectively meaningful. They nevertheless simplify the problem of assessing diverse or multidimensional effects (whether subjective or objective) with one another, and help to reduce experimental error.

The first reference system was the European Master Telephone Reference System (SFERT). SFERT was adopted by the CCIF in June 1928. The equipment at the SFERT laboratories was a replica of the fundamental reference system then used in the United States of America. Reference equivalents were established by direct or indirect comparisons of the test connection with the reference system, based on loudness comparisons. A second reference system was developed called ARAEN based on the fundamental one metre air path. The concept of a fundamental one metre air path was specified and developed by the UK Post Office, and adopted by the CCITT in 1948. Articulation ratings (AEN) were established by direct or indirect comparisons of the test connection with the reference system, based on articulation. ARAEN stands for "Reference apparatus for the determination of articulation ratings". In 1960 due to the unavailability of essential components it had become impossible to maintain the SFERT reference system. Hence a series of tests was carried out to determine what changes to the ARAEN system were required to render it suitable as a direct replacement for the SFERT system. The new reference system was called NOSFER (New master system for determining reference equivalents)

The transmission quality of reference systems are much higher than that of an ordinary handset telephone. Loudness ratings for commercial telephones connections are determined by direct subject comparisons with a reference system. The results provide useful indications of the effective overall mouth to ear losses. However reference systems are not practicable, for making more complex comparisons, were the effects of noise and frequency distortions are to be included. The inclusion of noise and frequency distortions are essential if the results are to be used for planning telephone networks.

2.2.2 Standard Speech Link (working standard)

For telephone network planning, it is usual to use a `transmission standard`. The use of the word ` standard` can be ambiguous in this case, a more appropriate term to refer to such a connection is a `representative limiting connection` RLC. for short. This is a telephone connection made up of telephone equipment representative of the apparatus in service. The connection consists of two local telephone circuits connected by a fixed junction attenuator. The whole system is set up to represent the worst connection that can occur in a national telephone network. If a RCL is used with an adjustable attenuator for tests the equipment maybe termed a `standard speech link` or a `representative telephone connection`. This equipment is used for determining communication efficiency due to practical factors of real telephones connections. The equipment can be divided into two `ends` and used separately , so that such properties as loudness ratings can be measured separately for the send and receive directions of any local telephone circuit under test. A `standard speech link` can be set up to include the distortions that could occur in real telephone connections, such as circuit noise additional attenuation / frequency distortion.

However there was a problem of specification due to the variation of commercial microphones and earphones. This was overcome by the use of a Intermediate Reference System (IRS).

2.2.3 Intermediate Reference System (IRS)

An IRS represents a telephone connection which contains calibrated handsets. The handsets have linear microphones and earphones, which are quite stable and can be calibrated quite accurately. An IRS can be used as a `standard speech link`, when used with suitable circuit components, including amplification. The greater stability of the microphones and earphones avoids many of the problems encountered in assembling basic telephone planning data, such as those due to manufacturing tolerances and other variations in commercial telephone microphones and earphones.

2.2.4 Measurement of Reference Equivalents and Relative Equivalents

Measurement of true reference equivalent and relative equivalents are known as telephometric measurements. To determine the "true" reference equivalent, the test connection is compared by voice and ear directly with the NOSFER reference system, ref.(29). However it is not usual to make direct comparisons with the NOSFER system, generally only working standards are compared directly. The reference equivalents of working standards are determined before they go into service and from time to time afterwards. Hence most reference equivalent measurements, are made from comparisons with working standard systems. These measurements are referred to as "measurement of relative equivalent".

2.3 Speech Assessment Methods

2.3.0 General

Any meaningful definition of speech "quality" must be based on human responses and perception. Objective quality measures on the other hand, attempt to measure those physical attributes of the speech signal that are correlated with factors which determine speech quality. Since an objective test measures success and is generally evaluated by it`s ability to predict some subjective quality assessment. The performance of an objective measure cannot be dissociated from the subjective measure it estimates. Subjective measures can be broadly grouped into two categories Utilitarian and analytic. Utilitarian methods have three goals;

1 - To be reasonably efficient in test administration and data analysis.

2 - Measure speech quality on a unidimensional scale.

3 - The test method must have good efficiency, reliability and repeatability.

The most important characteristic of Utilitarian methods is that the results are expressed on a unidimensional scale. Thus results are summarized by a single number, enabling direct comparison of the relative quality of speech communication systems. Tests used to assess the communication efficiency of telephone connections are utiliarian . In contrast analytic methods seek to identify the underlying psychological components of the speech that determine perceived quality. Also, to discover the acoustic correlates of those components. The results of such experiments are usually expressed in more than one dimension and are orientated toward characterizing speech perception rather than measuring perceived quality.

2.3.1 Various Speech Assessment Methods

The invention of the telephone necessitated the development of speech quality tests for quantifying the communication efficiency of telephone connections at different listening levels due to the variance in line attenuation of different connections. A method for assessing the quality of a voice channel in terms of it`s articulation value was introduced. Articulation was defined as the percentage of transmitted speech units that were correctly identified by the listener. The speech units could be phonemes, monosyllables, words or sentences. Obviously when comparing different speech channels, the same speech units must be used, because word and sentence perception are affected by semantic and syntactic factors. Fletcher`s and Stainberg`s methodology formed the basis of two kinds of speech tests;

(i) Intelligibility tests, which are scored by the percentage of correct identification of the meaning conveyed by the transmitted speech.

(ii) Articulation tests, which are scored by the percentage of correctly identified sounds, words or sentences in the transmitted speech.

Articulation tests can be modified so that scores have a comparable origin by the introduction of a reference speech system, see section (2.2). The listening levels of the test and the reference speech channels are reduced in a controlled manner by increasing the line attenuation. The score is expressed in terms of "articulation equivalent loss" (AEN), which is defined as the difference between the two attenuation values that give the reference and test speech channels 80% articulation. An example of articulation value as a function of attenuation is plotted for a reference and test channel is shown in figure 2.3. Figure 2.3 shows the "articulation equivalent loss" (AEN) for the test system, i.e. the difference in attenuation applied to the test and reference systems that produces a articulation score of 80%.

Measurement of AEN

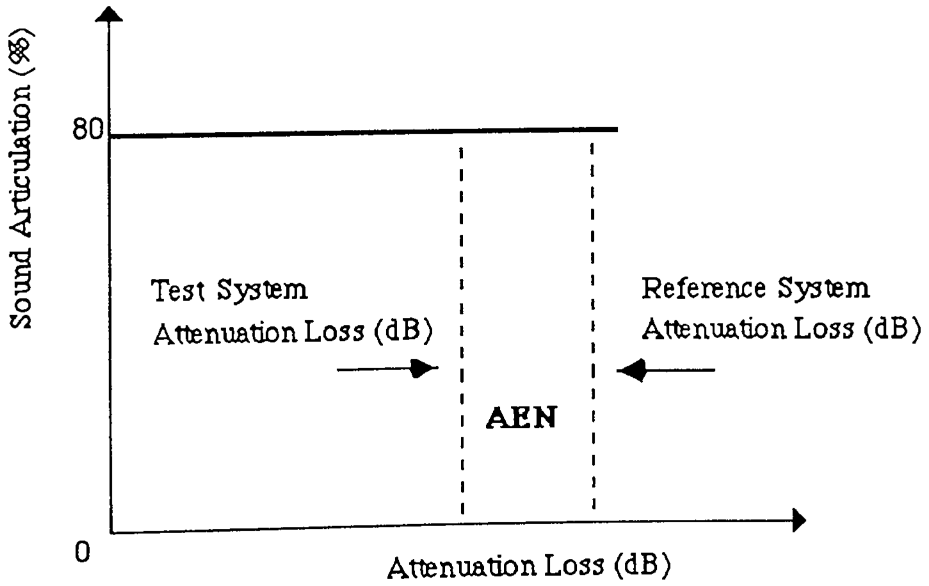


Figure (2.3)

2.3.2 Rhyme Tests

The articulation test was modified by Fairbanks (7), the listener responses were limited to a set of "rhyming" words. The test consisted of C-V-C monosyllables. Listener response sheets had the trailing vowel consonant specified, and listeners only had to identify the leading consonant from the test speech. This method allowed the intelligibility of single phonemes to be measured.

2.3.3 Modified Rhyme Test (MRT)

House modified the rhyme test by limiting the listener responses to a set of six rhyming words. This approach also improved the administration of the tests. The test lists had fifty words, hence the response sheets consisted of fifty sets of six rhyming words. New test lists could be made by randomly selecting one word from the six on the response sheet, new response sheets could be made by randomizing both the order of the sets and the order of the words in the sets. This type of test was more suitable for use with untrained listeners because subjects had to identify words instead of nonsense monosyllables. Also the use of a finite message set effectively eliminated the variation due to word frequency effects.

2.3.4 Diagnostic Rhyme Test (DRT)

Voiers ref.(36) changed the possible listener responses so that they differ not only in the leading consonants but also to restrict them to differ in just one distinctive feature of the leading phoneme. The listener response sets are limited to two words, one of which is the stimulus word. The distinctive features considered in the DRT are voicing, nasality, sustension, sibilation, graveness and compactness. DRT tests not only yield an overall intelligibility rating for a speech system, but also a "diagnostic" rating consisting of intelligibility scores in each of the distinctive feature categories.

2.3.5 Speech Interference Test

Nakatani proposed a speech quality test based on articulation principles, but with application to speech quality measurement. The procedure involves introducing interfering speech into a reference system and the system under test. The interfering speech introduced into each system is the same as the speech transmitted by the system, (i.e. vocoder, low pass filtered, etc). Nonsense sentences are used to reduce semantic effects, the two systems were scored as percentage articulation. A fixed level of interfering speech is added to the reference and test systems. The test results are used to identify two thresholds, T-0 and T, which are the speech-to-noise levels associated with 50 percent articulation for the reference and test signals, respectively. This is illustrated in in Figure (2.3). The quality of the test speech is denoted as Q and is measured in dB, and is the difference between the thresholds T-0 and T.

2.3.6 Speech Quality Tests

Speech intelligibility tests are only suitable for communication systems that produce moderate to severely degraded speech, because of their very nature they are unable to differentiate between two speech systems that are highly intelligible. Highly intelligible systems may differ in other perceivable attributes such as pleasantness, naturalness or required effort to understand the speech. Tests that differentiate between highly intelligible speech systems are more appropriately termed speech quality tests.

2.3.7 Pair Comparison Methods

The "isopreference" method ref.(17) uses pairs of test signals each having a different speech level and degraded by different levels of additive noise are passed through the transmission system under test. Subjects listen to all pair combinations the test signals (N test signals = (N(N-1))/2) and specify which they would prefer to use for a telephone call. The subjective results are plotted by means of "isopreference contours" in the two -dimensional parameter space of speech level verse noise level. Each test signal is plotted via it`s speech and noise levels. Test signals that had equal preference are joined by a line, (isopreference contour). Using this method contours explicitly indicate the optimum speech transmission level for a system`s expected background noise level. To convert the isopreference contours to a unidimensional scale i.e. "Transmission Preference Scale" (TPL) or "Transmission Preference Units" (TPU) . The test system is compared to a reference system at its optimum level. One of the first methods for quantifying the performance of telephone systems, had the from of a paired comparison test. Loudness Rating`s used reference equivalents based on perceived loudness.

2.3.8 Loudness Rating

Loudness rating tests aim to provide a measure of the transmission loss of a speech system from the mouth of the talker to the ear of the listener. The result is a unidimensional measure related to the loudness with which the listener perceives speech emitted by the talker at a specified constant level. The principle on which the tests are based, involves comparing the system under test with a reference system NOSFER, to establish a "reference equivalent". The full procedure specifying the equipment and method can be found in ref (29).

2.3.9 Direct Methods for Measuring Subjective Speech Quality

A limitation of all paired-comparison preference / equality tests is that the reference signals can only represent a limited range of speech distortions. Since the subjects are required to judge the test speech in terms of the reference speech, listeners may be limited to a smaller perceptual descriptor space than might be desired. The most widely used direct method of subjective quality measure is the category judgement method, which results in a mean opinion score (MOS). In this method listeners rate the speech under test on the following five-point scale.

Rating	Speech Quality	Level of Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible, but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying , but not objectionable
1	Unsatisfacttory	Very annoying and objectionable

Only five categories may seem to few, studies on the limits of human information processing indicate that as few as five but no more than nine categories should be used. The mean of the subjects opinion scores is taken as the measure of perceived speech quality. The test method consists of two parts, the first is a training session where subjects listen to three reference signals which represent the high, middle, and low quality range of the speech they will hear. This process in meant to give all the listeners the same subjective range and origin in their quality ratings. A standard set of reference signals must be used if the results are to be compared to the results of other test sessions carried out at different times and places.

2.3.10 Listening Effort Tests

Subjects listen to small groups of non-related sentences, each group of sentences are heard at one of five different listening levels.After each sentence group subjects express their opinion on the following scale.

Opinions Based on the effort required to understand the meaning of sentences

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

The responses are scored A=4, B=3, C=2, D=1, E=0 these scores are statistically analyzed as functions of the properties of the speech. The Listening Effort test procedure and analysis are described in more detail in section (4.5).

2.3.11 Indirect Judgement Tests

Indirect judgement tests are designed to measure the underlying parameters that determine perceived speech quality, these methods are used in an attempt to reduce preference differences among individual listeners. The procedure involves listeners judging the test speech on several "parametric quality scales" rather than one overall quality / Acceptability scale. The early work in the area of parametric descriptors of speech quality was done by Voiers ref.(35) and McGee ref.(18). The method was developed by Voiers in his Paired Acceptability Rating Method (PARM) and in the Diagnostic Acceptability Method (DAM). Examples of parametric scale descriptors are "clicking or ticking," "babbling or gurgling," and "fluttering or twitering." These scales provide an indirect measure of quality / acceptability by measuring attributes of the of the speech that determine quality / acceptability . The principle behind indirect judgement methods is that subjects generally agree on the degree to which a speech degradation is present in the test speech, but vary in their preference of that specific degradation.. An estimate of a systems overall quality / acceptability are derived from the parametric quality scores.

2.3.12 Communication Efficiency Tests

Conversation Tests are used for assessing the communication efficiency of telephone connections, which is inherently a measure of speech quality. This involves the determination of "message rate efficiency" by measuring task performance. This concept has many variations, the idea is for the subjects to perform a specified task that requires communicating over the system under test. One method proposed by British Telecom (formerly British Telephone Administration) in 1959 and now incorporated in CCITT recommendations. Requires subjects to carry-on a two-way conversation over a telephone connection as a means of accomplishing the task. An example of the type of task, is for one talker to describe a simple line drawing to the subject at the other end of the connection. The object is to complete the task in as short a time as possible. The subjects are encouraged to engage in conversation to clarify the task immediately after their task the subjects to rate the difficulty of their task in terms of conversational effort, based on the following five point scale.

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

The line attenuation is increased up to a point where the articulation scores on both systems are substantially reduced. Sound articulation verse attenuation graphs are plotted. Using the resulting curves, a value of 80% sound articulation is taken to compare the attenuation values of the test system A1 and the SREAN system A2. Articulation reference equivalent (AEN) is equal to the difference between the two attenuations A1 and A2.

$$AEN = A2-A1$$

Chapter 3

Literature Search of Problems of Assessing Synthetic Speech

3 Assessment of Synthetic Speech

3.0 General

The assessment of synthetic speech would seem relatively straight forward, considering the amount of experience gained from the assessment of natural speech over a telephone connection or radio link. It could be assumed that the techniques developed for speech assessment over the telephone or radio could be directly applied to synthetic speech. However, caution should be exercised because it has been shown that in some areas special problems arise in assessing synthetic speech. This section is devoted to a survey of some areas where special problems arise in assessing synthetic speech and a review of previous published work on the assessment of synthetic speech.

3.1 The need for reliable assessment

Reliable methods of assessing the quality and acceptability of speech output from synthesis systems would be useful to two groups ; (a) designers of the hardware and software and (b) users of voice output products. Designers of the hardware and software require diagnostic tests so that modifications to the system can be systematically quantified. Potential users of synthesizers are currently able to choose among a fairly small range of commercially available products, but this will increase as the use of voice output systems becomes more obvious. At some point it will become necessary to agree on a method for judging which will be the most suitable device for a specific task. Researchers designing these new products need a method for assessing which are the best amongst existing devices with a view to determining why they are good, and ideally will be able to measure their own designs against a standard.

Although it is not difficult to see why it is necessary for assessment, it is not a simple matter to agree on techniques of evaluation. A number of questions can be asked, should techniques be subjective or objective - or a combination of both . Additionally how much weight should be placed on where the synthesizer will actually be used? For example the conditions over a telephone line will be different from a air plane cockpit, and yet again different in a classroom. Is it essential to measure performance on specific tasks? Another factor to consider is the difference in the way synthetic and natural speech are perceived. This factor is fundamental to answering the above questions. Mainly because of the limitations imposed on listeners short term memory by synthetic speech. Before discussing the differences between synthetic and natural speech perception it is useful to reiterate the fundamental processes of speech perception.

3.2 Perception of speech in general

It is generally accepted that the perception of speech is a parallel interactive process which identifies speech using acoustic-phoneme information and linguistic knowledge sources. This is often referred to as a bottom-up top-down process sec. (1.3.4). The degree to which a listener relies on each source is dependant on the quality and context of the speech. For example the perception of high quality spoken sentences will use a greater amount of linguistic knowledge and less acoustic-phoneme information compared to the perception of single words. The difference will be dependant on the predictability of the word in the sentence. Which is based on the available linguistic information, context, semantic and syntactic information. Another factor to be considered is the specific perceptual strategies adopted by listeners. Specific cognition processes used are imposed by the redundancy of the acoustic cues and listening conditions.

3.3 Perception of Synthetic and Natural Speech

Previous research has demonstrated that synthetic speech is harder to understand than natural speech ref.(26) Several hypotheses have been suggested as to the reason why. The most popular are that synthetic speech is equivalent to a form of degraded natural speech, (i.e.. natural speech in noise), an error in timing (i.e. the time between phonetic cues is incorrect or unnatural ref.(10) or as suggested by Pisoni (26) the acoustic-phonetic cues are impoverished in some way. It has already been stated that listeners adapt their cognitive identification processes to cope with differing amounts of linguistic and acoustic information contained in speech. For synthetic speech they must adapt their perceptual and identification strategies to a signal which by its nature is different to natural speech. It could be concluded that the poor performance of synthetic speech corresponds to a combination of the quality of the speech and specific cognitive processes adopted by listeners. It therefore follows that the performance of synthetic speech could be improved by "tuning" subject's perceptual and identification strategies by training.

3.4 Perceptual Spaces and the Identification of Natural and Synthetic Sentences

N. Bacri ref.(3) used an intelligibility gain experiment to investigate the question of whether the acoustic-phonetic structure of synthetic speech was equivalent to degraded natural speech, or had an impoverished acoustic structure. The study used 8 phonetically balanced lists of 10 sentences. Each sentence list used one of several syntactic structures. Sentences were reproduced using natural speech, L.P.C. speech, and two synthesis by diphones text-to-speech systems. The stimuli were intensity balanced and degraded using masking pink noise. Prosodic and phonetic cue effects were strong, while the effect of syntax was weak. The procedure involved adjusting the signal to noise ratio (SNR) so that none of the stimuli were intelligible. The (SNR) was then increased in stages and after each change subjects attempted to identify as much of the stimuli as possible. This method enabled the researchers to identify differences in the perceptual strategies adopted by subjects for different stimuli.

The results of the study showed that degradation of speech by pink noise varied mainly as a function of the speech systems. In general the subjects perceptual processing of the speech relied mainly on acoustic-phonetic cues, and secondly on prosodic segmentation cues. The presence of backward lexical identification mechanisms provided evidence of the use of higher order information. However this was dependant on the main effect, i.e.the quality of the speech produced by each system. Bacri concluded that these results agreed with Nusbaum and Pisoni's conclusion ref.(23) "the differences in perception of natural and synthetic speech are largely the result of differences in the acoustic-phonetic structure of the signals". Bacri also concludes that the dissymmetry between the natural and synthetic speech responses support the hypothesis that synthetic speech is "impoverished speech". Different generating systems (natural, LPC. and synthetic speech) produced different patterns of cues to the listeners. Which required listeners to construct and process several "perceptual spaces".

3.5 Effects of Training on Perceptual and Identification Strategies

A study of perceptual learning carried out by Greenspan et al (8) investigated the effects of training on test results. The results showed that the training of subjects by exposure to specific synthesized material in general improved the subjective scores in recognition tests. In Greenspan's study three groups of subjects were used, one group were trained using synthesized isolated words, another group used synthesized sentences, and a control group of untrained subjects. The results showed that the word-trained subjects showed improvement in isolated-word verification tasks compared to the control group. But no corresponding improvement was found in sentence verification tasks. In contrast, the sentence-trained subjects improved on both the isolated-word and sentence verification tasks.

These results were interpreted as " the sentence-trained subjects were acquiring special strategies for segmenting words in fluent speech that were not acquired by the word-trained subjects". The adoption of specific perceptual strategies impose different amounts of cognitive loading on listeners Short Term Memory (STM.). A suggestion by Simpson et al, (31) that the increased cognitive load associated with synthetic speech is due to the unfamiliar accent of the speech. If this was true it would be expected that Greenspan's study would have shown an equal learning effect for words and sentences.

3.6 Cognitive Loading in the Processing of Synthetic speech.

The differences in perceptual strategies adopted by listeners in subjective tests, have implications for transferring the results of subjective evaluation to a decision about what is the best synthesizer or best research plan for improvements. For example some speech may require perceptual strategies that impose greater demands on listeners short term memory (STM.). Listeners may not be aware of the increased demand or "cognitive loading" on their STM. But listeners exposed to this type of speech for long periods of time, or are carrying out other tasks at the same time, may experience greater fatigue than normal. Obviously the effects of cognitive loading will influence where such speech can be implemented.

3.7 Word Processing Time Tests

One effect of increased cognitive loading is an increase in the processing time in identification tasks. Nusbaum and Pisoni (21) carried out several studies which involved tests on the processing time for words and non-words produced by synthetic and natural speech. A lexical decision task was used to compare the time required to classify a stimulus as either a 'word' or 'nonword'. It was found that the mean response times for natural words and nonwords (903ms & 1046ms) were significantly faster than for synthetic words and nonwords (1056ms & 1179ms) (Slowiacczek & Pisoni 1982 ref.(32)). Experiments carried out by Luce et al., 1983 ref.(15) which involved the recalling of digits and words serial order. Showed that more short-term memory (STM) was required for processing synthetic speech than for natural speech.

From these results it would appear that the phonological encoding of synthetic speech requires more cognitive effort (increased cognitive load) compared to the perception of natural speech.

Nusbaum and Pisoni (1984) ref.(23) came to the conclusion that " the problems in perception of synthetic speech are largely tied to the processes that encode and recognize the acoustic-phonetic structure of words". In order to compare the redundancy of synthetic and natural speech. Manous and Pisoni (16) used a word gating task, which measured the amount of stimuli required before a word was identified. The redundancy of the synthetic speech was less than that of natural speech, i.e. synthetic speech required a greater proportion of the stimulus before a word was identified. The degree of redundancy of speech, has implications for where the speech can be used. If the redundancy of the speech is low, it would be unsuitable for use in a noisy environment.

3.8 Speech Quality

There is no general accepted definition of vocal quality, it may be referred to as the total auditory impression the listener experiences upon hearing the speech of another talker or speech output device. Speech quality could be referred to as a subjective measure of a combination of several attributes of the speech, attributes such as naturalness, pleasantness, intelligibility, and distinctness. Generally the speech quality of a rule synthesizers were initially judged in terms of intelligibility at the phoneme and word level.

Some text-to-speech synthesizers are capable generating highly intelligible speech . The fact that the speech is intelligible does not mean that the speech will sound natural or be acceptable to the general public. For example if every word is over-articulated and spoken at a very slow rate, with long pauses between words, then the speech will be intelligible, but at the same time totally unacceptable. For the synthesized speech to be acceptable, other attributes of the speech such as naturalness, pleasantness etc. which contribute to the overall perceived speech quality have to be considered.

3.9 Measurement of Speech Attributes

An example of a study designed to measure the attributes of reproduced speech by suitable subjective scaling techniques,was reported in a paper "Subjective assessment of automatic voice answering machines" ref. (6). The paper presents results which demonstrate how the profiles of suitable attributes illustrate important differences in the perceived characteristics of five voices. The five voices were made up of, natural speech and four different implementations of speech concatenated from stored segments of speech encoded by linear predictive coding. Subjects used four semantic differential scales sec. (4.4.2) Listening Effort, Acceptability, Pleasantness, and Naturalness to rate each voice. The scales were scored 0 - 4, i.e. 4 on the naturalness scale indicated the listener perceived the voice as natural. ('Listening Effort' was scored 4 for minimum effort and 0 for maximum effort). The four scores represent a profile for the relevant voice. The profile for voice one (natural speech) differed markedly from those of the other (synthesized) voices. Interesting distinctions can also be seen between the profiles of the four synthesized voices. A feature common to the synthesized voices is that the listening effort scores are appreciably higher than the naturalness scores. Indicating that a lack of listening effort (more intelligible) can be partially separated from the naturalness of the speech.

3.10 Segmental and Suprasegmental Levels of Evaluation

Segmental evaluation involves the evaluation of phonemes and words, where as suprasegmental evaluation is concerned with sentences and paragraphs, and the affects of prosody. The majority of assessment work on the quality of rule synthesized speech has been at the segmental level. This is because most of the work involved diagnostic tests on synthesis systems which were still under development. Many tests from other areas, like speech audiometry and testing of analogue or digital communication channels were used to quantify the development systems. Tests like the Modified Rhyme Test (MRT), the Diagnostic-Rhyme Test (DRT), and the use of phonetically -balanced (PB) CVC words have been used to measure the segmental intelligibility of synthetic speech. A study carried out by Pols & Olive (1983) ref.(27) provides evidence of a potential problem in using MRT and the DRT in their original form. The dyadic rule-synthesis system and the reference LPC -system tested in the study, showed many b-v confusions, whereas the response alternative v - b rarely occurs in the standard MRT with 6 response alternatives.(Nye & Gailarby 1973) ref.(25).

To make it easier to run the MRT and DRT tests meaningful words were used for stimuli as well as for alternative responses. This makes it easier to run tests with naive subjects without training. However, this could influence the percentage correct scores and the confusions matrices. If a phoneme is not correctly identified, resulting in a nonsense

CVC word, then the subject might guess the correct word again since he realizes that meaningless words cannot occur. If the tests were being used for a diagnostic evaluation of a synthesizer it might be better to use an open response test with nonsense words.

3.11 Segmental Evaluation of Several Synthesis Systems

Logan, Pisoni and Greene ref.(14) carried out an experiment which used several variations of the standard Modified Rhyme Test (MRT. sec. 2.3.3) to measure the segmental intelligibility of synthetic and natural speech. Eight text-to-speech systems were tested along with a control (reference) natural speech system. The experiment examined the reliability of the MRT when used with synthesized speech. The results showed that overall error rates for the synthetic speech were higher than for natural speech. It was noted that if the error rates of initial consonants only were compared one synthetic voice produced error rates as low as the natural speech. Significant differences for overall error rates were found between many of the synthesizer systems. However several similarities also emerged, patterns of phoneme errors were observed that were common for different voices. Logan suggests that this is due to problems in the phonemic implementation rules used for segmental synthesis. It is worth noting that at the time of this study most of the synthesis systems used were still under development. A possible explanation for the similarities in the phoneme error patterns, is that several or all of the synthesis systems tested used phonemic implementation rules originally based on Klatt's pioneering work . Klatt produced a comprehensive set of rules for the automatic conversion of American English text into phonemes.

3.12 Listening Effort in Intelligibility Tests

When comparing the intelligibility of synthesizers, the results may show two synthesizers are not statistically different when the test stimuli are short sentences. If one system is slightly less intelligible than the other system. The results may not be significantly different because the listeners used more effort to understand the less intelligible speech. This may not be noticeable in the short term, but if the same systems intelligible speech. This may not be noticeable in the short term, but if the same systems were used for longer more linguistically complex stimuli. The listener will be more fatigued due to the greater effort required to understand the second system using extended stimuli. This emphasizes the need to consider what the synthesizer is going to be used for. Both systems maybe suitable for automatic information services, train times automatic banking, etc where the output speech is usually concise short sentences. Whereas the second system would be unsuitable for applications such as the speech output device for a reading machine for the blind.

3.13 Intelligibility of synthetic CVC stimuli over the telephone

Malsheen et. al. carried out an experiment to measure the effects of telephone bandwidth on the segmental intelligibility of two commercial text-to-speech systems derived from the M.I.Talk synthesis system.sec.(1.5). An open response rhyme test was used with monosyllabic CVC English words to test the robustness of acoustic cues for consonants generated by rule. The phonemic error patterns were compared to see if there were any similarities between the two systems. The two systems were tested under two conditions, with and without telephone simulation. Results of the study showed that the intelligibility of both systems were reduced when the bandwidth was limited. However, the segmental degradation patterns differed between the two systems. To explain the difference between systems Malsheen suggests that the primary and secondary cues for particular phonemes were weighted differently. This difference in weighting may also account for the differences that are observed in the perception of natural and synthetic speech. That is, a cue which is normally present in natural speech may be present to a greater or lesser extent in synthetic speech. If the relationship between phonemic-cues in synthetic speech was already tenuous, it would be further degraded by telephone bandlimiting. The results of this study confirm Pisoni's hypothesis of synthetic speech having "impoverished" phonetic cues. Malsheen et al conclude that, "What may be "impoverished " is the natural balance between primary and secondary cues for particular segments".

3.14 Intelligibility of synthetic CVC stimuli in noise

Pratt (25) carried out a similar study designed to quantify the intelligibility of 7 commercially available text-to-speech synthesis systems. The work was part of a study examining the suitability of speech as a input / output channel for Army Battlefield Data Communication systems. The experiment used a Diagnostic Rhyme Test (DRT) (sec. 2.3.4) to measure the intelligibility of single words. Subjects recorded their opinion using a rhyming word pair that differed by a single acoustic attribute in the initial consonant. Six attributes were used: voicing, nasality, sustention, sibulation, graveness and compactness. The voices were tested under two listening conditions clear, and with added speech like noise (SNR. 0 dB). Analysis of the DRT. results revealed two main sources of variance, firstly listening conditions and secondly voice source system. In clear conditions some of the synthesis systems scores were not significantly different from the natural speech scores. But under noisy listening conditions the natural speech score was significantly higher than all of the synthesis systems. The degree of degradation of the synthetic voices in noisy listening conditions was system dependant. Which was seen as a change in rank order of the synthetic speech systems. After each DRT. subjects rated speech they had just heard on four semantic differential scales Naturalness, Pleasantness, Intelligibility and Effort (required to comprehend). The results from the rating scales were correlated against each other and the DRT score. Generally the rating scales were highly correlated amongst themselves except for the naturalness scale. Pratt also noted there was a slightly higher degree of correlation between the DRT scores and and the Effort rating scale than the DRT and intelligibility rating scale.

3.15 Testing Overall Speech Quality

It has already been mentioned that subjective assessment of synthesized speech has on the whole been based on intelligibility tests. These tests were used for diagnostic purposes assessing the performance of text-to-speech systems which were still under development. The tests measured the intelligibility of specific linguistic units phonemes, words, and sentences. However such tests do not provide any information of a listener's reaction or preference to a particular synthesizer. As the intelligibility of synthetic speech approaches that of natural speech. The utility of a particular text-to-speech system will become increasingly dependant on the speech quality. Which is based on the attributes of the speech, such as naturalness, pleasantness, acceptability etc. Specifically speech that is irritating, tiring, or boring will be perceived to be of lower quality by listeners than speech that is pleasant and natural sounding.

3.16 Trust or degree of confidence.

In 1983 Nusbaum, Schwab and Pisoni reported the results of initial tests aimed at obtaining systematic judgments of the subjective quality of natural and synthetic speech, ref (24). Nusbaum refers to subjects perception or impressions of particular attributes of speech as "subjective reactions". For example one subjective reaction may concern a listener's impression of the naturalness of the speech. Another subjective measure of speech quality is the trust ref. (24) or degree of confidence listeners would place in the speech of a particular talker (either natural or synthetic). This response could reflect the confidence listeners have in their ability to understand the talker, as well as some less well defined concept of reaction to the talker. The first factor is clearly related to the intelligibility of a particular talker, which is independent of whether the speech was produced by a human or a synthesizer. While the second factor is probably related to the listener and the circumstances under which the speech is heard. For example a person who has had little or no exposure to synthesized speech may distrust speech that sounds as if it has been generated by a computer. Alternatively a person who is familiar with computers may have greater trust in computer generated speech than recorded natural speech.

3.17 Human Factors:- Preference

The need to determine whether one synthesis system is more, or less, suitable for a particular application than other synthesis systems is not just one of determining which system has the highest intelligibility score. But whether one system is more acceptable or preferable to users. Acceptability or preference measurements of one type of speech over another for a specific task.are inherently measures of speech quality. Preference tests have been used in the past to measure the preference for and the acceptability of natural speech under different transmission or degradation conditions. One example is the isopreference method ref.(17), were subjects made preference judgments on speech samples at various combinations of signal to noise ratios relative to a standard reference condition. There are several basic criticisms of this approach, mainly it is not clear how to generalize this method, i.e. yield curves of equal preference in a space of signal to noise ratio, to the subjective evaluation of synthetic speech. Bacri's work on perceptual spaces ref. (3) provides evidence of differences in the perceptual processing of natural speech in noise and synthetic speech. Which casts doubt on the validity of isopreference measurements based on stimuli that require different perceptual strategies. As the isopreference method depends on the feasibility of constructing a set of reference samples of speech that are equally spaced in terms of their perceived quality.

Other methods of measuring the perceived quality of speech have included the use of a rating scale or magnitude estimation, and multidimensional scaling of judgements (McGee, 1964, 1965) ref.(18) along with Osgood's (1952)* semantic differential technique. The major drawback to the rating scale approach as well as magnitude estimation and category judgments is that only a single measure of quality along a single scale is elicited with no further information about other subjective impressions or attributes. However Voiers (1977) ref.(35) has generalized the rating scale approach in the Diagnostic Acceptability Measure in which listeners rate speech samples on several different rating scales, with each scale intended to measure a different perceptual quality (e.g. "raspiness"). Similarly, the semantic differential approach elicits judgments of speech quality on a number of different scales defined by opposing adjectives (e.g., "annoying"/"pleasant").

Other comparative tests have been used to measure the acceptability of speech systems.Voiers ref.(35) compared the performance of several vocoders in different noise conditions using the Diagnostic Rhyme Test of intelligibility and the Diagnostic Acceptability Measure. He concluded that the acceptability of coded speech is strongly related to the intelligibility of the speech. Simpson and Marichonda-Frost (1984) ref.(31) used semantic differential scales to measure the effects of speech rate and pitch on the acceptability of synthetic speech generated by the the Votrax ML-1 synthesizer. Their results indicated that the rated acceptability of the speech varied as a function of speech rate, even though intelligibility was not affected by rate manipulations. These studies illustrate, measurements of speech quality and acceptability have only been used in very constrained ways to evaluate synthetic speech.

3.18 Evaluation Tests for evaluating synthetic speech

3.18.1 Comprehension & Semantic Differential Tests

Pisoni carried out a study (26) to develop an evaluation test. To be used for systematic tests that would indicate important subjective differences between natural, and samples of speech produced by different text-to-speech systems. In his study the speech was evaluated using a comprehension test based on information contained in the test speech, and a semantic differential test (17-scales). Subjects also had to estimate how much trust they would place in different kinds of information provided in the speech. Finally subjects answered questions on their reactions to certain aspects of the experiment, which included estimating their performance in the comprehension test.

The study concluded that it is possible to systematically measure listeners perception of the naturalness of speech. The differences between the natural and synthetic speech that appeared to be unrelated to intelligibility focused on suprasegmental qualities of the speech. Synthetic speech was judged to be more coarse, choppy, old, harsh, rough, and foreign than natural speech. These qualities were related to the general prosodic characteristics of glottal source, intonation, and timing information. Thus, for the judgments related to the signal the relative naturalness of the speech can be separated to some extent from the intelligibility of the speech ref.(37). Another conclusion :- The acceptability of synthetic speech appears closely related to the segmental intelligibility of the speech (cf. Voiers, 1980) ref.(37) The measurement of acceptability is based on subjective estimates of trust or confidence in different categories of information. Acceptability may depend more on the segmental intelligibility rather than the naturalness of the speech. The results of the study suggest that for most subjects intelligibility rather than naturalness is more important when considering the acceptability of speech. In addition, based on the subject's estimates of their performance in the comprehension task, it appears that the confidence of a user listening to synthetic speech is also a function of the intelligibility of the speech. It was also noted, that even though there were minimal differences in actual comprehension performance for the different types of speech, subjects who heard synthetic speech underestimated their performance.

3.18.2 Diagnostic Rhyme Test & Semantic Differential Test

Pratt used a Diagnostic Rhyme Test (DRT.) to quantify text-to-speech synthesizer performance. The speech material comprised 96 rhyming word pairs that differed by a single acoustic attribute in the initial consonant. There were six attributes; voicing, nasality, sustention, sibulation, graveness and compactness. Paid subjects who had experience listening to degraded natural speech, but had no experience with synthetic speech were used. Twelve voices were evaluated, eleven synthetic and a natural control made up of six male voices. The twelve voices were assessed under two signal-noise ratios. An analysis of variance of the results revealed that there was an interaction between listening conditions and voices. Which indicates that different voices DRT. scores degraded by varying degrees. This was seen as a change in the rank ordering of the synthetic voices. The natural voice had the smallest change in DRT. score indicating it was the most robust. These results are in good agreement with Malsheen et. al.ref. (19)

results. Another significant interaction was between initial consonant attribute and signal-to-noise ratio (listening condition). The attributes of nasality and sibilation were least affected, with graveness showing the largest reduction in score.

Subjects also took part in two semantic differential scaling tests. The first test was run in conjunction with the DRT test, hence the results are based on word stimuli. Before each session subjects were played a series of ten second samples of all the synthesizers in both clear and noisy conditions. This was done so that listeners could hear the range of speech quality to be used in the experiment. This was done to encourage the listeners to use the full range of the rating scales. One draw back of conditioning listeners in this way is that the test systems are rated in a relative sense, rather than an absolute sense with respect to the scales. Therefore the results are considerably dependant on the stimuli used. Results showed that subjects were strongly influenced by the added noise and were unable to differentiate between systems. In contrast the DRT. became more discriminating in the presence of noise. The scores from the semantic differential scales were correlated with each other and the DRT. scores. There was a high degree of correlation amongst the rating scales, particularly for intelligibility, Effort, and pleasantness. Pratt suggests that the subjects appear to regard these semantic terms as highly synonymous. The correlation between DRT. and intelligibility (r=-0.88) was slightly less (but not significantly) than the correlation between DRT. and effort (r=-0.90).

The second semantic differential scaling test used the same scales as above but used sentence material as stimuli. Three different passages of speech were used so that the effects of speech material could be assessed. The experimental design was such that the any material effect was separated from a possible learning effect. Analysis of the results involved a factor analysis of the four scales. The first factor accounted for 81% of the variance. Pratt suggests that the subjects perceptual space represented by the four semantic scales, is essentially unidimensional and may be approximated by combining the results from the four scales.

3.18.3 Acceptability Field Test & Semantic Differential Test

The two evaluation tests were used to assess the subjective speech quality of a synthesis-by-rule weather information system, ref. (34) The test system was the commercially available text-to-speech synthesizer Infovox SA 101. Twenty two telephone subscribers took part in a field test designed to measure the acceptability and a number of quality parameters of the synthetic speech. Subscribers called a weather service where the messages were synthesized. Two 5-grade annoyance scales, Pronunciation & Stress were used to measure the acceptability of the speech. Plus six semantic differential scales (SDS) were used to measure attributes of the speech. The semantic differential scales were Intelligibility, distinctness, comprehensibility (effort to understand meaning), naturalness and pleasantness.

The second test, carried out under laboratory conditions used 25 phonetically balanced sentences presented alternately by natural and synthetic speech. Twenty subjects rated the quality of the synthetic speech in relation to the natural speech, (ratio estimation). Responses were recorded on a scale 0 -100% where 100% was defined as being the quality of the natural speech. Four scales were used, Overall quality, Intelligibility, Naturalness and Pleasantness. Subjects were familiarized with synthesized speech by giving them a page of literary prose text to read while the corresponding text-to-speech synthesis was replayed to them.

The natural voices were assessed using a paired comparison test. So that the differences between the natural voices could be measured and taken into account. The perceived performance of the synthetic speech was summed up as follows; " the occurrence of incorrect pronunciation and incorrect stress is rather annoying, while the syllabic rate is just slightly too fast. Intelligibility and comprehensibility is fairly good, but the sound is rather unpleasant and unnatural. The overall quality is on average judged fairly good, although with some greater weight on lower ratings".

Chapter 4

Work Carried Out and Experimental Methods

4. Methods and Tests used in the present study

4.0 General

The work carried out in this study was for British Telecom who were developing a reference system to be used for the assessment of text-to-speech synthesis systems. Reference devices have been used in the past to measure the quality of speech over a telephone connection in terms of specific properties of the connection. None of the reference devices previously used seem particularly suitable for use with synthesized speech now being introduced into public telephone networks. It is desirable for the speech produced by a reference device to be similar to the speech it is being used to assess. Generally the choice of reference device is governed by the form of the predominant degradation and the types of perceptual errors incurred. Hence an adjustable attenuator was used for assessing transmission loss, added speech interference for sidetone, and speech modulated noise was used for quantization effects. Therefore it may be assumed that the resulting perceptual errors produced by the reference device and the test speech will be similar. 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. Although comparisons can be made between systems which sound very different, the listener`s task is made much easier if they do sound similar. It cannot be claimed that Time Frequency Warped (TFW) modulated speech sounds exactly like any particular version of synthesized speech, but it can be said that the resulting displacements in timing do have some resemblance to errors produced by synthesized speech. A further consideration is the degradation of synthetic speech by the telephone connection. It has been shown that synthetic speech degrades to a greater degree compared to natural speech under the same conditions ref.(7). This is due to the synthesized speech being less redundant than natural speech ref(26). Another factor to consider is the degree of degradation of synthetic speech is system dependant ref (28). Hence it is important to assess individual synthesis systems rather than synthesis techniques such as, Text-to-speech, allophone, diphone, analysis resynthesis etc. British

British Telecom laboratories Martleshem Heath have proposed a new reference degradation sec. (2.0), Time Frequency Warping (TFW) modulation, for assessing synthesized speech.

4.1 Time Frequency Warped Modulation (TFW)

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". The effect was produced by sampling real speech at a rate well above the Nyquist rate (8KHz) and storing it in a computer (PDP 11). The samples were then reproduced 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 (squarewave mod.) on a sinewave is shown in figure (4.1).

Time Frequency Warping Modulation of a Sinewave

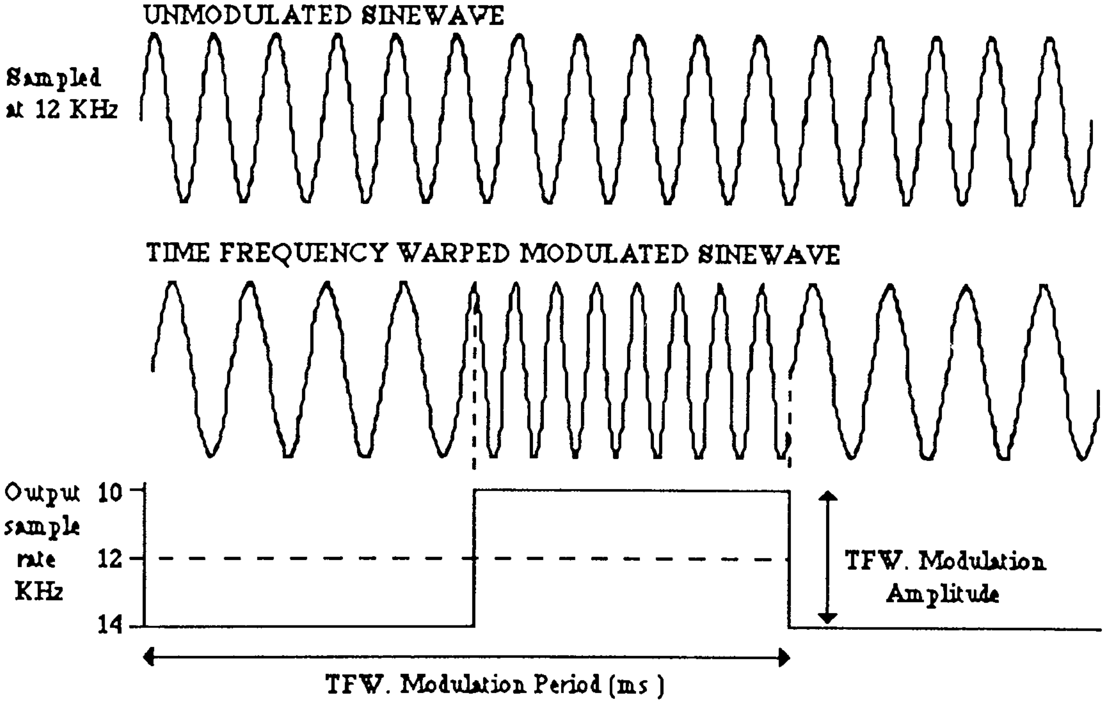

UNMODULATED SINEWAVE
Sampled at 12 KHz
TIME FREQUENCY WARPED MODULATED SINEWAVE
Output sample rate KHz
10
12
14
TFW. Modulation Amplitude
TFW. Modulation Period (ms)

Figure (4.1)

TFW modulation has three parameters for varying the output sampling rate, modulation waveform, period, and amplitude. It is desirable for a reference device 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.

4.2 Aims of this Research

This study assesses the suitability of TFW. modulation for use as a reference device for measuring the quality of synthesized speech. This involved determining the extent to which TFW modulated speech was perceived by subjects to have some similarity to synthetic speech. Secondly to determine which parameter of the TFW hardware should be used as the variable parameter, to provide the assessment scale. Thirdly to make some assessments in terms of that scale of typical examples of synthesized speech.

4.3 Assessment Methods Used

4.3.0 General

At present there is no standard assessment method or criteria for rating the quality of speech produced by speech synthesis systems. The assessment of speech synthesizers to date has been mainly in terms of diagnostic, articulation and intelligibility tests. Diagnostic tests such as the modified rhyme test MRT sec.(2.3.3) have been used to quantify the performance of text-to- speech synthesizers at the word and phoneme level. Articulation and intelligibility tests have been used to quantify the overall intelligibility of synthesized speech. TFW modulation is being developed for use as a reference device to be used in conjunction with an assessment test to provide a standard assessment method for quantifying the performance of text-to-speech synthesizers over a telephone connection. The assessment test has to be able to differentiate between synthesis systems of similar or completely different quality. Articulation and intelligibility tests are sensitive to small changes in intelligibility up to about 70% intelligibility, but above this figure the sensitivity decreases fig. (4.2).

Assessment of Telephone Connexions According to Various Criteria

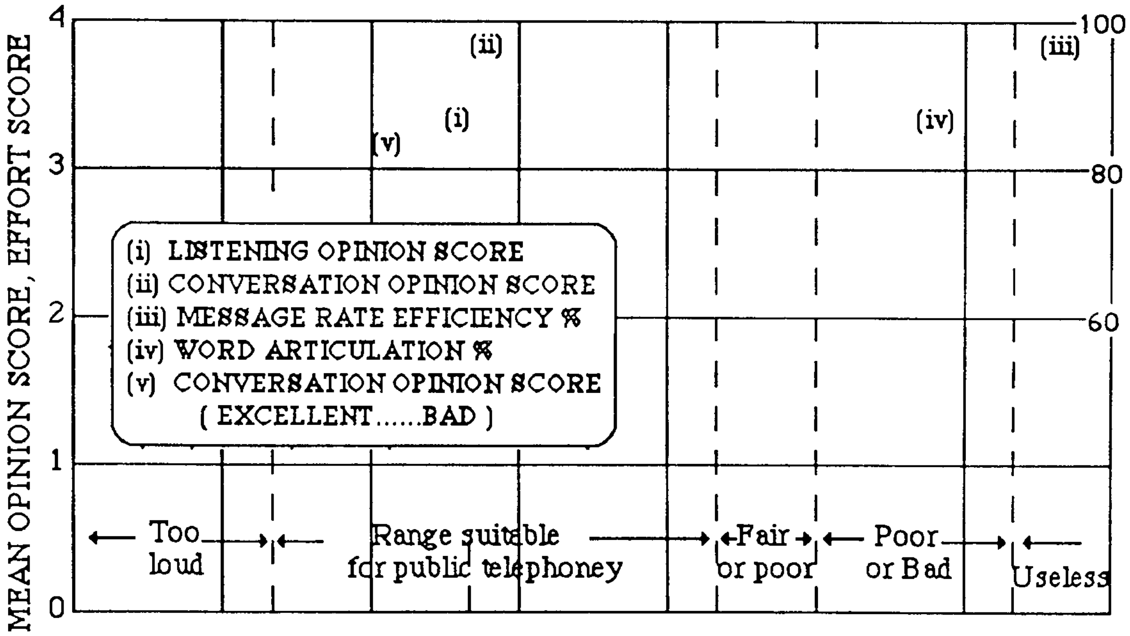


Figure (4.2)

Therefore, for highly intelligible synthetic speech i.e. 95 - 100%, the ability of such tests to differentiate between high quality synthesizers is limited. For synthesizers to be used over the telephone, it is important that the quality / intelligibility of the synthesized speech is high, because the speech will be degraded to some degree by the telephone link. It is desirable to have an assessment method that can differentiate between highly intelligible text-to-speech systems, so that the best possible system can be found. A further problem with articulation and intelligibility tests is that the results do not indicate the amount of effort required to understand the speech. This is important because, if a synthesizer yields a high intelligibility score but requires a lot of effort on the part of the listeners to achieve this score. It would render the synthesizer unsuitable for use over a telephone link. Because any degradation of the speech introduced by the telephone link or the user's listening conditions would make the speech more difficult or impossible to understand. To differentiate between systems with similar intelligibility scores a listening effort test can be used. Listening effort tests measure the effort required to understand the "meaning" of sentence stimuli. The proposed standardized test for quantifying the performance of text-to-speech synthesis system over a telephone connection uses the TFW modulation in conjunction with a listening effort test.

4.3.1 Listening opinion test.

Listening effort tests have been used in the past by telephone engineers to measure the "quality" of telephone links. Intuitively it would seem reasonable to use such tests to quantify the performance of synthesizers that are proposed for use over the telephone network. Such tests were used in the past to quantify the performance of digital coding devices and other sources of degradation over a telephone link. Listening effort tests have proved to be reliable, repeatable and are able to differentiate between small changes of speech quality. The test does not attempt to define speech quality in any absolute way, but provides a comparison between two or more candidate systems. If a reference device is used in the test, it will provide a standard by which the other speech systems can be rated

.

4.3.2 Analytic Measures

Part of this study is to identify the attributes of speech subjects use in determining the perceived difference between different types of speech. For example, are subjects basing their opinions of speech quality on the naturalness, intelligibility or combinations of attributes of the speech. Determining the perceived similarity between TFW modulated and synthetic speech involves determining the speech attributes used by subjects to differentiate between the two types of speech. Semantic differential scales have been used in tests to differentiate between different speech synthesis systems. The tests measure specific attributes of speech. Comparison of different synthesis systems were then made based on the attributes measured. This approach compares speech synthesis systems based on the rank order of the systems on the particular semantic differential scales, i.e. naturalness. In this study the semantic differential scaling approach was modified to determine the relationship between the attributes measured, and the test subjects perception of the relationship between the test speech systems.

Multidimensional scaling (MDS) tests were used in conjunction with semantic differential scaling (SDS) tests to determine the perceived relationship between TFW modulated, synthesized and natural speech samples. The combined use of MDS and SDS tests provides a method of determining the perceptual relationship between different speech samples. They also determine the attributes of the speech samples on which the relationship is based.

Multidimensional scaling and semantic differential scaling are analytical methods, which were used in an attempt to characterize the underlying psychological factors that determine perceived speech quality. As well as providing an insight into speech quality perception, multidimensional and semantic differential scaling were used as practical subjective tests for measuring speech quality and determining the perceptual relationship between different types of speech. The combined multidimensional scaling and semantic differential scaling tests used in this study combine uni and multi dimensional scales to assess the perceived difference between speech samples.

4.3.3 Uni and multi dimensional scales.

A unidimensional scale measures one specific attribute or characteristic of the speech under test, i.e. naturalness or intelligibility. Whereas multidimensional scales measure a non specific quantity such as the difference between two speech treatments. When using unidimensional scales in subjective tests the subjects are given a criteria on which to base their opinions. For subjective tests using multidimensional scales the criteria given to the subjects is non specific. Hence the attributes of the speech on which the subject's opinions are based are unknown to the experimenter. The subjects themselves are probably not aware of the exact attributes they are basing their decisions on. The multidimensional scaling and the unidimensional semantic differential scaling experiments attempt to reveal which attributes of speech the subjects are using in differentiating between synthesized, TFW modulated and natural speech.

4.3.4 Consistency of Subjects Opinions

An inherent problem of subjective tests is the variability of subjective scores. For speech quality assessment tests this can be attributed to subjects having a varied perspective of what is the "ideal speech quality". Another source of subjective variance is the ability of the individual to give consistent opinions. This is particularly important in pair comparison tests where the criteria for making a judgement is non-specific. For example in the multidimensional scaling tests where the criteria for assessment is based on the perceived "difference" between two speech treatments. The effects of subjective variance can be reduced by randomizing the design of the experiment and rejecting subjects results which are considered inconsistent. In the multidimensional scaling tests, subjects rate the "difference" of a pair of treatments. It was found in preliminary work for the this study, that some subjects gave inconsistent opinions of the perceived difference between treatments. For example if a subject rated speech treatments S1 & S2 as the same, and S2 & S3 as the same, it could be assumed that S1 and S3 would be rated as similar. If the subject had rated S1 and S3 as completely different, it would be impossible to map these three points in multidimensional Euclidean space, where the distances between points corresponds to the perceived difference data. The difference between the distance and perceived difference data is measured by the "stress" (error) for the two sets of data, this is explained in appendix (7.1.1). If the stress is above the acceptable limit of 0.2 the data is rejected. It could be argued that a improved transform function would reduce the stress. But if the rest of the subjective data produced configurations that were acceptable, then the data with unacceptable "stress" should be rejected.

4.4 Determination of TFW. modulation variable parameter.

4.4.0 General

The initial part of this study was to reduce the three variable parameters of TFW. modulation to one, by fixing the settings of two of them. The variable parameters are modulation waveform, period and amplitude sec. (4.1). The first approach was to determine which modulation waveform would be the most suitable. A pair comparison experiment was carried out to compare the three TFW. modulation waveform settings. Each TFW. modulation waveform sinusoidal, square and triangular using the same range of period and amplitude settings were compared to a synthetic speech sample.

4.4.1 Simple Pair Comparison Experiment

Four combinations of Period and Amplitude were taken and, for each combination, three different waveforms were used, namely: Sine, Triangular and Square, which gave twelve combinations (treatments) altogether. The twelve TFW. treatments were each directly compared with a sample of synthesized speech. The object of the experiment was to obtain an indication which TFW modulation waveform produced TFW modulated natural speech, that would be perceived to be similar to synthesized speech.

Time Frequency Warping Modulation settings used.

Period ms	200	200	300	300
Amplitude KHz	2	3	2	3

(Note Amplitude in KHz refers to maximum variation from original sampling rate.) These combinations were chosen to provide speech roughly comparable in intelligibility to that of the synthesized speech. It may be remarked that, in the light of later study, the speech samples were of comparable quality, although this was rather poor by telephone standards.

Seven subjects compared each of the twelve TFW treatments, with a single example (treatment) of synthesized speech. The synthetic speech was produced by a text-to-speech software package run on a PDP.11 computer and output through a British Telecom synthesizer. Each treatment used the same pair of sentences;

"You will find me at home tomorrow"

immediately followed by

"What do you make of this"

The speech was presented to the subjects at a comfortable listening level on a handset by means of tape recordings via a simulated local area telephone system containing a 300-3400 Hz bandpass filter. The members of each pair of treatments were recorded four times in succession in parallel on separate channels of the recorder. A short burst of 1KHz tone was inserted after each treatment. The subjects listened to a treatment once on one channel and at the tone switched to the other; thus, they heard the two treatments being compared twice each and then gave their opinion. To assist the subjects a light was used to indicate when to score. The order of presentation of the treatment pairs was randomized. figure (4.3) shows the switching procedure followed by the subjects.

Controlled switching method used by subjects for comparing two speech treatments.

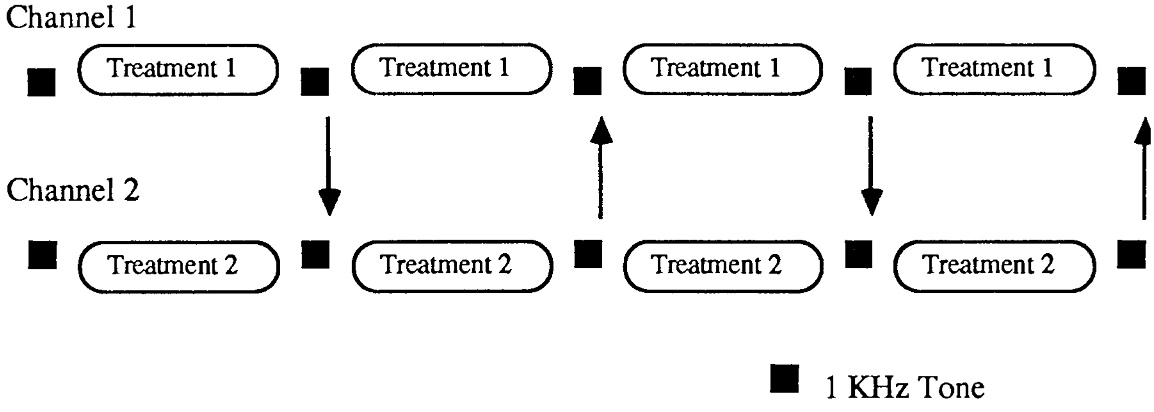


Figure (4.3)

This controlled method of switching between channels / treatments was used to avoid the problem of subjects switching within a treatment, and thereby finding that the speech samples were not synchronized in time. The fact the recording of the speech samples were not synchronized could have been interpreted as a difference between the treatments, rather than a difference between the timing of the recordings. Scoring was intended to measure the extent to which the subjects perceived difference between the two treatments. This was done by the use of the following "Semantic Differential" scale.

SAME --------------------------/----------------------- COMPLETELY DIFFERENT

Subjects indicated their scores by placing a mark at what they considered a suitable point along the scale. Their opinion was recorded by measuring the distance of the mark along the line, the values were used as in the analysis as a numerical value ranging from SAME = 0 to COMPLETELY DIFFERENT = 10. This method of scoring provides a measure of the perceived difference between the two treatments.The results are shown in appendix (7.2.1) It has already been stated that the quality of the synthesized speech used in the pair comparison test was of very poor quality. Probably because the text-to-speech analysis software used in the test was at the time still under development. The fact that the quality of the synthetic and TFW. modulated speech were very poor, and that the two types of speech did not sound particularly similar. Raised the question of, which attributes or characteristics of the speech, subjects were using to differentiate between the speech samples (treatments). Were subjects basing their decision of the difference between two treatments on one or more attributes of the speech? This question is investigated in section (4.4.2).

4.4.2 Multidimensional (MDS) and Semantic Differential Scaling (SDS) Experiment

Multidimensional and semantic differential scaling experiments were used to determine the attributes of speech subjects used to differentiate between TFW modulated, synthesized, and natural speech treatments. The output of a multidimensional scaling analysis is a table of coordinates which locate each object (speech treatment) in multidimensional Euclidean space. The distance between each object in the configuration corresponds to or is a function of the proximity (Perceived difference) associated with the two objects. Semantic differential scaling tests measure specific attributes of the speech treatments, i.e. "naturalness", the results were used to interpret the MDS. configuration. Two experiments were carried out to investigate the attributes of speech subjects used to differentiate between different types of speech. Each experiment used a combination of multidimensional and semantic differential scaling tests. The MDS part of the experiment uses a non specific scale i.e. a "difference" scale to measure the relative perceived differences between speech treatments. The SDS parts of the experiments specify the attributes of the speech to be measured by the use of particular semantic differential scales. The two experiments were carried out using the experimental procedure described below.

4.4.2.1 MDS & SDS Test Procedure.

The speech was presented to the subjects at a comfortable listening level on a handset by means of tape recordings via a simulated local telephone system containing a 300-3500 Hz bandpass filter.

In Part 1 subjects heard all of the treatments presented in random order. The subjects were required only to listen and familiarize themselves with the range of speech quality. This helped reduced the incidences where subjects want to change their score for a particular treatment they had heard earlier in the experiment after they had heard another treatment presented later.

Part 2 Measurement of Four Specific Attributes of the Speech Treatments

This part of the experiment used a semantic differential scaling test to subjectively measure attributes of the treatments. Subjects heard each treatment repeated twice, after each treatment repetition they scored their opinion on a number of semantic differential scales.

Part 3 Measurement of Perceived Difference between Treatments (MDS).

Subjects were presented with pairs of treatments and were required to give their opinion of the perceived difference between the two treatments. This was repeated for all the different pairs of the treatments, i.e. (N(N-1))/2 pairs. Multidimensional scaling analysis was used to transform the perceived difference data into a configuration in multidimensional space sec.(7.2.1).

Part 4 was a repeat of the tests in part 2

(Note:- The same group of subjects was used for all four parts of the experiment. The order of presentation of treatments / treatment pairs was randomized for all four parts of the experiment to avoid sequence affects)

4.4.2.2 Analysis of MDS and SDS Results

The computer program used to carry out the multidimensional scaling analysis for this study, was developed by D.L.Richards.Appendix (7.1.1.1). Principal component analysis was used to analyze the semantic differential scales data. Response data may have hidden factors which generate the dependence or variation in the responses. The observed responses can then be represented as functions of the latent factors or components. The mathematical form of these components must be one which will generate the covariances or correlations between the responses. If the number of hidden components is less than the number of initial variables then a simplified description of the dependence structure can be obtained. The principal component analysis used in this work, treats the speech treatments and their observations as linear components of the latent variables. The analysis of the dependence structure amounts to the statistical estimation of the coefficients of the functions. Principal component analysis is one of the simplest forms of factor analysis, the theory and computional procedure used in this study are described in appendix (7.1.2).

4.4.3 Listening Effort Test to Determine TFW Variable Parameter

The next stage of the study was to use a listening effort test to determine which of the remaining variable parameters period and amplitude of the TFW. modulation was to be fixed. Listening effort tests are based a unidimension scale measuring the effort required to understand the meaning of the test speech. The listening effort scale is not as specific as semantic differential scales which measure particular attributes of the test speech. But it is more specific than the "difference scale" used in the multidimensional scaling tests. Although the criteria for assessment of the test speech is specified in listening effort tests. In general the listening opinion scores will be based on more than one attribute of the test speech.

4.4.3.1 Listening Effort Test Experimental Procedure

The subjects were postgraduate and undergraduate students who answered an advertisement. Subjects were paid for taking part in the experiment. The sentence lists were made up of simple, short sentences, all different, and chosen at random as being easy to understand. There were 20 sentences in a list (5 groups of 4) with no obvious connection of meaning between one sentence and the next. The speech treatments were recorded on a Revox reel to reel tape recorder a separate tape was made for each subject according to the experimental design. The experiment is based on a randomized 11 * 11 graeco-latin square.The graeco-latin square used for the listening effort test used to determine the TFW modulation variable parameter is shown below.

Graeco - Latin Square

T2D	S1Y	NaB	T!E	T7F	S2J	T4H	T5G	T3I	S3A	T6C
S3C	T3H	T7E	S1I	T1Y	T2A	S2D	NaF	T4J	T6G	T5B
T5E	S2C	S1J	T4A	T3D	T6F	S3B	T1H	T2G	NaY	T7H
TIA	T6I	S2F	S3Y	T2E	T7D	NaJ	T4B	T5H	S1C	T3G
S2H	T1F	T5C	T7B	NaG	T4I	T3Y	T6A	S1E	T2J	S3D
T7J	S3E	T4G	T2F	S2B	NaH	T5I	T3C	T6Y	T1D	S1A
T6B	T4D	T1I	T3J	S1H	S3G	T2C	T7Y	S2A	T5F	NaE
S1G	T5J	T2Y	T6H	S3I	T1C	T7A	S2E	NaD	T3B	T4F
T3F	NaA	S3H	T5D	T6J	S1B	T1G	T2I	T7C	T4E	S2Y
T4Y	T7G	T6D	NaC	T5A	T3E	S1F	S3J	T1B	S2I	T2H
NaI	T2B	T3A	S2G	T4C	T5Y	T6E	S1D	S3F	T7H	T1J

Rows - represent subjects

Columns - represents sequence (order of presentation)

Letter - represent sentence lists

Letter / number - represents speech treatments

The tapes were replayed to the appropriate subjects via a variable attenuator and the listening end of a Intermediate Reference System IRS.) sec.(2.23). Each cell of the graeco-latin square (sentence list / treatment combination) were divided into five groups of four sentences. The subjects heard the five groups of sentences at five different listening levels. For each cell the order of the five listening levels was randomized to avoid sequence effects. The experiment was controlled as far as the subjects were concerned by two lights, one indicated when the group of sentences were about to begin and the other indicated when to score their opinion. The subjects gave their opinion of the speech according to the scale shown below.

Listening Effort Scale

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

Numerical scores are allocated to the responses as follows;

A = 4, B = 3, C = 2, D = 1, E = 0

The analysis of variance evaluates the following effects and tests them for significance, the analysis is based on the design graeco latin square, rows (listeners), columns (sequence effects), letters (lists) numbers (speech treatments) and the interaction of listening levels with each of the other factors.

4.5 Comparison Listening Effort Tests Assessment of TFW. Modulated and Synthetic Speech

A series of listening effort tests, (Comparison listening effort tests) using a number of amplitude settings of TFW. modulation, several synthesis systems, and natural speech were carried out. The propose of this series of tests was to:-

Measure the consistency of TFW. modulation listening effort scores

Comparison of subject variance for different TFW settings

Comparison subject variance for synthetic and TFW mod. speech, and test the significance of any differences

The reason for examining the variance of subjects scores, is if the variability of opinions for TFW. modulated speech is similar to that of synthetic speech of similar listening effort score. This would indicate that the group of test subjects were able to rate the two types of speech with a similar degree of accuracy. If for example the variance for TFW. modulated was significantly smaller than that of synthetic speech . This would make it more difficult to accurately determine a TFW.equivalent.

4.5.1 Establishing TFW Equivalent Scores

A reference device is used to rate the performance of a system in terms of equivalent settings of the reference system. Time frequency warping (TFW) is under consideration by British Telecom for use as reference device sec.(2.20) for assessing speech synthesis systems. In this study examples of text-to-speech synthesis systems were used in listening effort tests, along with examples of TFW modulated speech. To determine whether consistent equivalent listening effort scores could be established between the synthetic and TFW modulated speech. Linear regression was used to establish equivalent scores between the two types of speech. The method is describe below using some of the test results.

The results from the second listening effort test, 1st replication are discussed here. The mean opinion scores (mos) for the TFW treatments were transformed using the logistic transformation equation:-

$$yt = 100 \times \left[5 + 0.5 \times \mathrm{Log}_e \left[\frac{mos}{4 - mos}\right]\right]$$

and were plotted against the TFW amplitude settings. The y -on- x regression line and it's 95% confidence limits were calculated figure (4.4).

Regression Graph for Comparison Listening Effort Test 1 Repetition 1

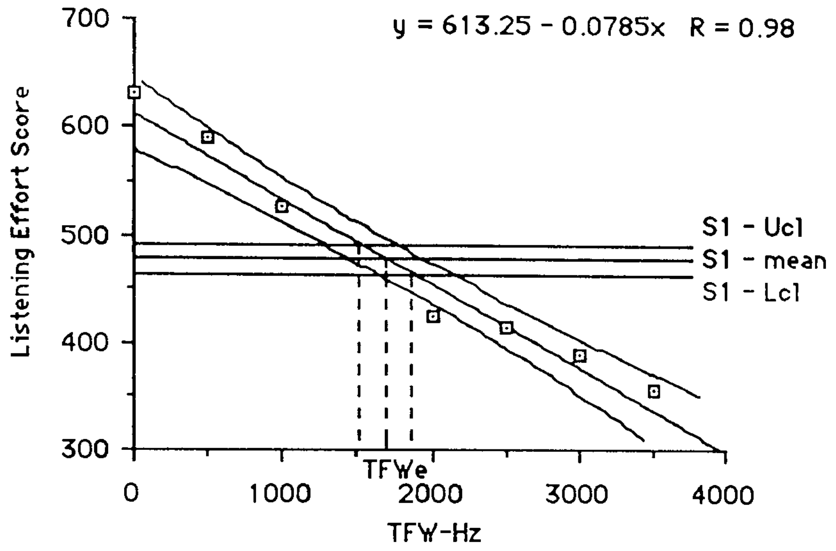


Figure (4.4)

In figure (4.4) the transformed values of synthesis system S1 and it's confidence limits are represented as straight lines. The mean opinion score for S1 was 1.58, which equals 479 when transformed. With upper (Ucl) and lower (Lcl) confidence limits of 493 and 463 respectively. The TFWe was calculated from the regression line equation. TFWe is the TFW modulation amplitude setting that would yield a mean listening effort score equal to the S1 synthetic speech treatment score. There is however a problem of estimating the accuracy of TFWe. This is due to the fact that the TFW and synthetic speech mean scores are subjective. Hence they have associated confidence limits which leads to a problem in assigning confidence limits to equivalent TFWe settings. To deal with this problem we assume that the regression line is the true or population regression line, hence it has no confidence limits.The intercept of the y -on- x regression line and S1 line is at x = 1707 (calculated from regression equation) is the TFW equivalent. The confidence limits for the TFWe equivalent are estimated from S1's confidence limits by calculating the intersection of the confidence limits with the regression line. These are shown in figure (4.4) as the two vertical broken lines. The confidence limits for TFWe were calculated from the regression line equation, the results were 1528 and 1910 Hz. What these figures mean is that 19 out of 20 times the TFW modulation setting that will produce a listening effort score equivalent to S1 lies within TFWe's confidence limits. With the most likely setting being TFWe. The confidence limits of the regression line at TFWe represent the 95% confidence interval of the expected mean listening effort score with the most likely value being the regression line value. It can be seen from figure (4.4) that the confidence intervals for the regression line at TFWe and S1 mean are similar. Hence this increases confidence in the mean listening effort score being the correct equivalent score.

Chapter 5

Results and Analysis

5. Test Results and Analysis

5.0 General

5.1 Simple Pair Comparison Experiment

The aim of the experiment was to determine which of the TFW modulation waveforms would produce speech that is perceived to be similar to synthesized speech. The experiment used each TFW. modulation waveform sinusoidal, square and triangular using the same range of period and amplitude settings. Eight subjects compared each TFW sample with an synthetic speech sample, using the experimental procedure described in sec. (4.4.1). For each TFW modulation waveform, four combinations of period and amplitude were used, i.e,

Period ms	200	200	300	300
Amplitude KHz	2	3	2	3

This gave twelve TFW modulated speech treatments, triangular (T1 - 4), sinusoidal (S1 - 4) and squarewave (R1 - R4).

5.1.1 Analysis of Results

The mean perceived differences between the TFW modulated speech treatments and the synthetic speech treatment were calculated. The results were plotted as a column graph Fig. (5.1) to give a simple visual comparison of the perceived differences.

Plot of The Mean of Subjects Perceived Difference Between TFW Modulated Natural Speech Treatments and Synthetic Speech Treatment.

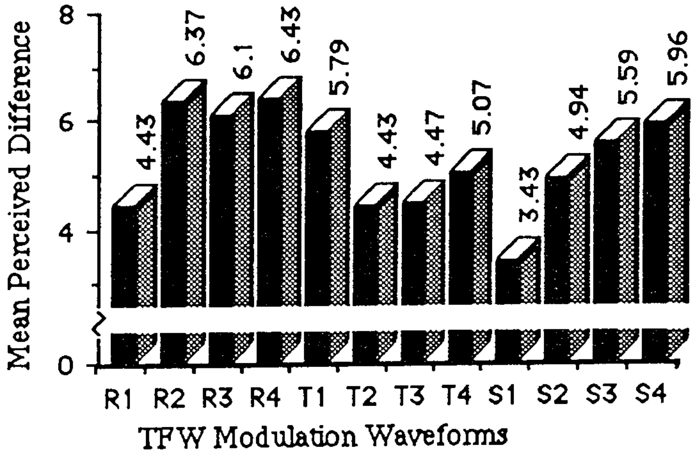
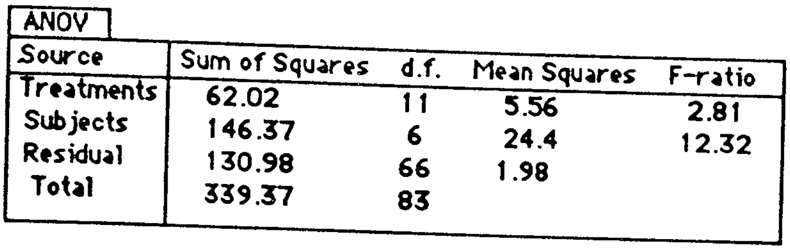


Figure (5.1)

An analysis of variance was carried out on the results to determine whether there was any significant difference between the treatment mean scores, i.e. between the mean perceived differences from the synthetic speech.

Analysis of Variance of Simple Pair Comparison Data



ANOV

Source	Sum of Squares	d.f.	Mean Squares	F-ratio
Treatments	62.02	11	5.56	2.81
Subjects	146.37	6	24.4	12.32
Residual	130.98	66	1.98	
Total	339.37	83		

Table (5.1)

The F-ratios for the subjects and treatments were compared with standard F-ratio tables to determine whether the observed effect was "significant". "Significant" means the level of the observed effect is such that there is a probability smaller than * of obtaining an equal or greater observed value by chance. The F-ratio from standard tables for two independent variables are shown below.

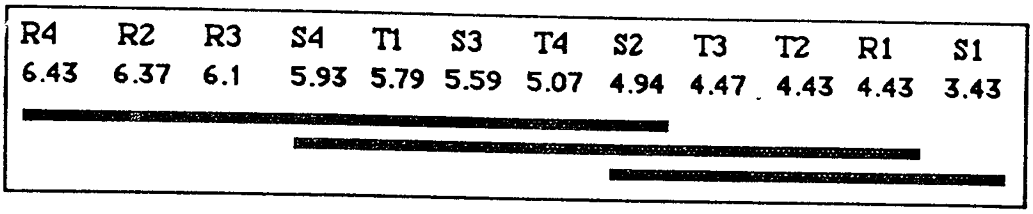
Level of Significance	* 0.05	0.025	0.01
F- (d.f. 11,66)	1.95	2.22	2.56
F- (d.f. 6,66)	2.25	2.63	3.12

Comparing the F-ratios form the analysis of variance table with the above figures, it can be seen that the variance between treatments, and between subjects are highly significant. The variance between subjects is as expected. The variance between treatments requires clarification. The treatments were placed in rank order and divided into "non-significant" groups, i.e. the difference between treatment means within the group are not statistically significant. To determine the groupings of the treatments the "least significant difference" was calculated.

$$\text{L.S.D.} = S\sqrt{\left(\frac{2}{n}\right)}\,t$$

Where S is the square root of the residual mean square, n is the number of observations of each treatment and t is the "student's t test" value for eleven degrees of freedom at the 95 percent confidence interval for the sample mean. The least significant distance for the treatment means was 1.51. Homogeneous groupings of treatments were calculated. Figure (5.2) shows the rank order of the treatment means and treatments that are not significantly different are indicated by a line drawn under them.

Rank Ordering and Grouping of the Speech Treatment Means Simple Comparison Test



R4	R2	R3	S4	T1	S3	T4	S2	T3	T2	R1	S1
6.43	6.37	6.1	5.93	5.79	5.59	5.07	4.94	4.47	4.43	4.43	3.43

Figure (5.2)

The mean and range of the subjective means for the four period and amplitude settings of the Time Frequency Warping were calculated for each waveform.

Rectagularwave -	Range = 2.00	Mean = 5.83
Sinewave -	Range = 2.5	Mean = 4.97
Triangularwave -	Range = 1.36	Mean = 4.96

The rectangular TFW modulation treatments produced the largest mean 5.03, which indicates that they were in general perceived as the most "different" when compared to the synthesized treatment. The mean of the treatment means for sinewave and triangularwave TFW modulations were virtually the same. The aim of the experiment was to obtain a indication of the TFW modulation waveform which would produce TFW modulated natural speech, which would be perceived to be similar to the synthetic treatment. If we look at fig (5.2) and the rank order of treatment means above, it can be seen that none of the waveforms stand out as a obvious choice. Sinewave TFW modulation was chosen because the four period/amplitude settings used in the experiment covered the largest range and contained the treatment that was perceived to be the least different from the synthetic treatment.

5.2 First MDS and SDS Experiment To Determine The Perceived Differences Between Seven Synthetic & One Natural Speech Treatment

The experiment compares eight treatments, six were produced using the B.T. text-to-speech synthesis system, which is a derivative of a system developed by the Joint Speech Research Unit JSRU. For each treatment the sentence input data to the synthesizer was either typed in as text or as a phonetic transcription of the text. Some of the treatments had modifications to the duration and pitch parameters of the phonemes. The other two treatments used were natural speech and speech synthesized using a unmodified JSRU synthesis system, details of the treatments are shown below. One of the aims of this experiment was to determine if simple modifications to the phoneme parameters would improve the quality of the synthetic speech. The experiment also attempts to identify the attributes of speech used by the subjects in differentiating between treatments. The dimensional / spatial relationship of the various synthetic speech treatments in cognitive space were investigated. Multidimensional and semantic differential scaling tests were used, these two tests are discussed in sec.(4.4.2). The treatments used in all four parts of the experiment are shown below.

S1 - PDP.11 Text input, no modifications

S2 - As S1 with 2 & 7 added respectively to duration & pitch parameters of phonemes

S3 - As S2 but with short pauses inserted between words

S4 - PDP.11 Phonetic input

S5 - As S4 with 2 & 7 added respectively to duration & pitch parameters of phonemes

S6 - As S5 but with short pauses inserted between words

S7 - JSRU Text-to- speech synthesizer.

S8 - Natural speech

The sentences used for all of the treatments were;

" He carried a bag of tennis balls."

Immediately followed by,

" It did not seem like summer."

These two sentences were used in both the MDS. and SDS tests because it was the difference between the types of speech we wanted the subjects to assess, and not the effect of sentence / speech combinations. The semantic differential scales used in this test were;

Intelligibility - Is it easy or hard to understand separate words in the speech

Naturalness - Does the speech sound natural or unnatural

Distinctness - does the speech sound clear or slurred

Pleasantness - Does the speech sound pleasant or unpleasant

Subjects were presented with each treatment repeated twice, and were required to give their opinion of the treatment based on the above attributes, on the scales below.

EASY -/---/- HARD

NATURAL -/---/- UNNATURAL

CLEAR -/------------------------------ ------------/- SLURRED

PLEASANT -/------------------------------ ----------/- UNPLEASANT

5.2.1 Measurement of Specific Attributes of Treatments. Results for Part 2 and Part 4 of the experiment

The numerical data for the subject's scores was produced by measuring the distance score along the semantic differential scales eg, Unnatural = 0, Natural = 10. The table (5.2) shows the mean subjective scores for the two S.D.S. tests, plus the mean scores of the two repetitions.

Treatment Mean subjective scores for Semantic Differential Scales MDS-SDS Test 1

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility R1	4.86	3.99	4.21	4.59	4.46	4.92	4.75	9.23
R2	5.33	4.61	4.07	6.42	5.34	4.74	5.28	10.05
Mean	5.10	4.30	4.14	5.51	4.90	4.83	5.02	9.64
Distinctness R1	3.74	3.17	3.32	4.47	3.17	4.08	3.39	9.04
R2	4.78	3.33	2.99	5.47	4.17	3.86	4.10	10.03
Mean	4.26	3.25	3.16	4.97	3.67	3.97	3.75	9.54
Naturalness R1	2.73	2.18	2.06	3.60	2.37	2.98	3.15	9.13
R2	3.77	2.41	2.33	4.31	3.06	2.64	2.96	9.85
Mean	3.25	2.30	2.20	3.96	2.72	2.81	3.06	9.49
Pleasantness R1	2.85	2.30	2.50	3.80	3.30	3.00	3.31	8.60
R2	3.67	2.52	2.48	4.53	3.41	3.09	2.96	9.71
Mean	3.26	2.41	2.49	4.17	3.36	3.05	3.14	9.16

Table (5.2)

From the above table it can be seen there are differences between the mean scores of the individual treatments. There are also a differences between the mean scores for a particular treatment over the two repetitions of the SDS. tests. To determine whether the difference between the first and second repetitions is significant i.e. statistically different, an analysis of variance sec.(5.1.1) was carried out. The result is shown
is shown in table (5.3).

Analysis of Variance of SDS Data Repetitions 1&2

ANOV - Reptitions					
Source or Variance	Sum of Squares	Degrees of Freedom	Mean Squared	F - ratio	Sig.
Replications Residual Total	25.46 4646.05 4671.51	1 278 279	25.46 16.71	1.52	NS

Table (5.3)

The analysis of variance showed that the results for the two repetitions were not significantly different. Therefore it was decided to use the mean of the two repetitions scores, for the rest of the analysis of the semantic differential scales. A attribute profile of each treatment can be produced based on the attributes measured on the semantic differential scales. Figure (5.3) shows the profiles of all of the treatments.

First MDS-SDS Test Treatment Profiles Mean of Repetitions 1 & 2

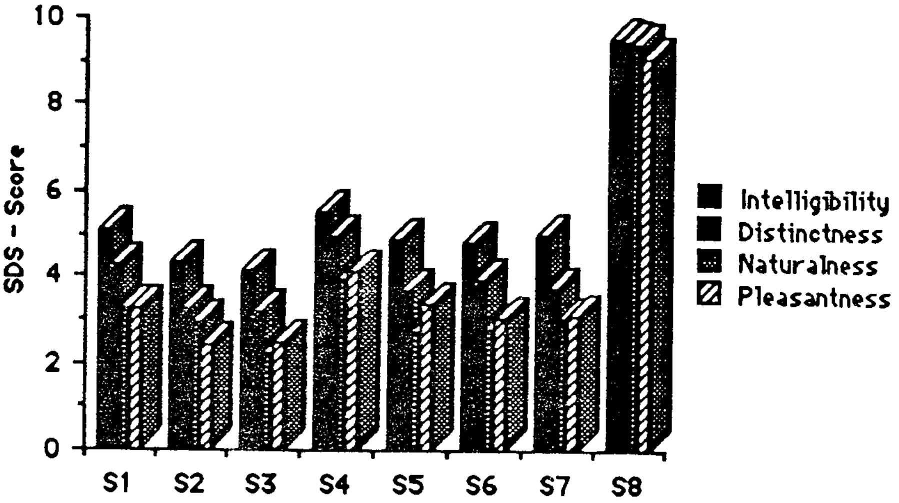


Figure (5.3)

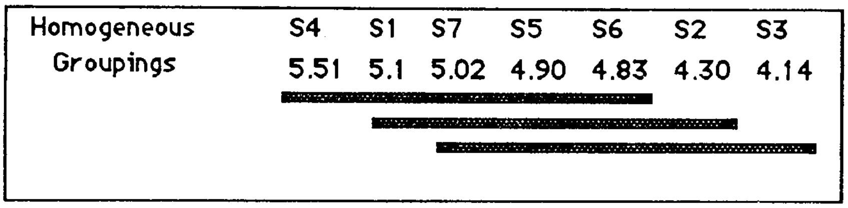
5. 2.2 Analysis of the Treatments SDS Mean Scores.

An analysis of variance was carried out for each of the four semantic differential scales. The analysis does not include the natural speech treatment because it would have a disproportionate effect on the variance. The analysis determines whether the differences between the treatment mean scores are significant. For each semantic differential scale the treatment means were placed in rank order. Using the calculated least significant difference between treatment means, the treatment means were divided into non significant groups. These groups are indicated by drawing a line under the treatment means within a particular group. The ANOVA tables and treatment groupings are shown below, Tables (5.4 a-d).

First MDS-SDS Test Homogeneous Groupings of Treatment Mean SDs Scores

ANOVA - Intelligibility - SDS

Source or Variance	Sum of Squares	Degrees of Freedom	Mean Squared	F - ratio	Sig.
Treatments	27.67	6	4.61	0.05	NS
Residual	1245.18	273	4.56		
Total	1272.83	279		L.S.D.	0.93

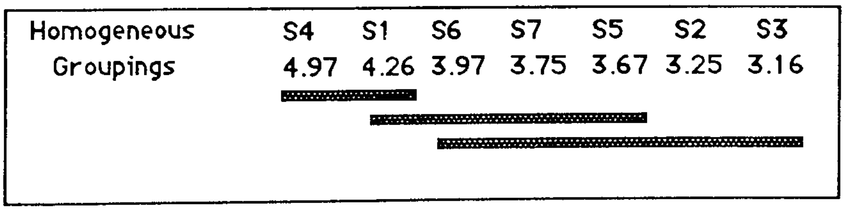


Homogeneous Groupings	S4	S1	S7	S5	S6	S2	S3
	5.51	5.1	5.02	4.90	4.83	4.30	4.14

Table (5.4a)

ANOVA - Distinctness - SDS

Source or Variance	Sum of Squares	Degrees of Freedom	Mean Squared	F - ratio	Sig.
Treatments	95	6	15.83	3.67	***
Residual	1178.4	273	4.32		
Total	1274.11	279		L.S.D.	0.9



Homogeneous Groupings	S4	S1	S6	S7	S5	S2	S3
	4.97	4.26	3.97	3.75	3.67	3.25	3.16

Table (5.4b)

ANOV - Naturalness - SDS

Source or Variance	Sum of Squares	Degrees of Freedom	Mean Squared	F - ratio	Sig.
Treatments	86.4	6	14.4	5.2	***
Residual	757.55	273	2.77		
Total	843.95	279		L.S.D.	0.72

Homogeneous Groupings	S4	S1	S7	S6	S5	S2	S3
	3.96	3.25	3.06	2.81	2.72	2.3	2.2
	━	━					
		━	━	━	━		
				━	━	━	━

Table (5.4c)

ANOV - Pleasantness - SDS

Source or Variance	Sum of Squares	Degrees of Freedom	Mean Squared	F - ratio	Sig.
Treatments	76.07	6	12.68	5.09	***
Residual	678.41	273	2.49		
Total	754.48	279		L.S.D.	0.68

Homogeneous Groupings	S4	S5	S1	S7	S6	S3	S2
	4.17	3.36	3.26	3.14	3.05	2.49	2.41
		━	━	━	━		
				━	━	━	
					━	━	━

Table (5.4d)

Using the L.S.D.diagrams, conclusions can be draw about the differences between the seven treatments based on the four attributes of speech used.

5.2.3 Determination of Perceived Differences Between modified and Unmodified Synthetic Speech Treatments

One of the aims of this experiment was to identify the speech attributes used by subjects to differentiate between different samples of speech (treatments). A multidimensional scaling test sec. (4.4.2.1) was used to obtain a relationship between the speech treatments based on the perceived difference. The result of the test is a configuration were the distances between the treatments approximate the perceived differences. The criteria on which subjects based their decisions of the difference between treatments is not specified and must be derived from the configuration. This involves identifying the attribute scales (and hence the dimensionality) on which the configuration is based. The results of the mean perceived difference between all treatment pairs are shown below as a "proximity matrix".

Proximity Matrix of Mean Subjective Scores

	S2	S3	S4	S5	S6	S7	S8
S1	6.05	6.05	0.98	5.33	4.60	2.31	9.06
S2		0.44	5.82	3.59	3.46	5.23	9.67
S3			5.98	4.15	3.11	7.28	9.61
S4				4.83	4.90	2.55	9.27
S5					0.71	6.63	9.33
S6						6.77	9.29
S7							9.50

The proximity matrix provides the input data to the multidimensional scaling analysis program app. (A.3.1). Which transforms the proximity data into distances between the treatments plotted in Euclidean space. The treatments are given a set of trial coordinates which are adjusted using a iterative process so that the distances between treatments approximates the perceived difference data. The resultant distance matrix is shown below;

	S2	S3	S4	S5	S6	S7	S8
S1	5.68	6.29	0.65	4.86	4.75	1.67	9.11
S2		0.65	5.75	3.18	2.63	6.79	8.33
S3			6.34	3.36	2.85	7.43	8.69
S4				4.56	4.53	2.22	9.71
S5					0.57	6.46	10.88
S6						6.29	10.35
S7							8.44

According to theory if the stress (goodness of fit) of the fitted distances is below 0.2 the fitted distances are said to be a good approximation of the perceived differences. The stress for the above fitted distances is 0.0195. The coordinates in Euclean space of the treatments from which the distance matrix was calculated were;

S1 - (-2.45, -1.64)	S5 - (-1.35, 3.1)
S2 - (1.72, 2.23)	S6 - (-0.84, 2.84)
S3 - (2, 2.81)	S7 - (-2.23, -3.3)
S4 - (-2.91, -1.18)	S8 - (6.06, -4.87)

A plot of the configuration is shown in figure (5.4), it should be noted that the orientation and scales of the axes is abitrary as long as the relative distances between the treatments is preserved.

MDS. Configuration of Differences Between Treatments First MDS-SDS Test

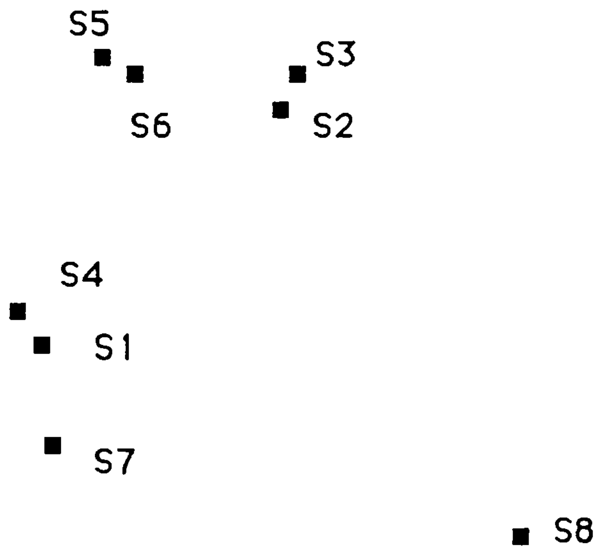


Figure (5.4)

It was stated earlier that the quality of the synthetic speech was very poor. This is reflected in figure (5.5), it can be seen that the synthetic speech treatments are clustered in two groups (S1, S4, S7) and (S3, S4, S5, S6) well away from the natural speech treatment, i.e. the synthetic treatments were perceived to be "very" or "completely" different, compared to the natural speech treatment. One of the objectives of this experiment was to determine if simple modifications to the phoneme parameters of the British Telecom synthesis system would improve the quality of the synthetic speech. This would be expected to be shown by the modified synthetic treatments (S2, S3, S5, S6) being grouped closer to the natural treatment than the unmodified synthetic treatments (S1, S4).

Perceived Differences of Modified Synthetic Speech Treatments

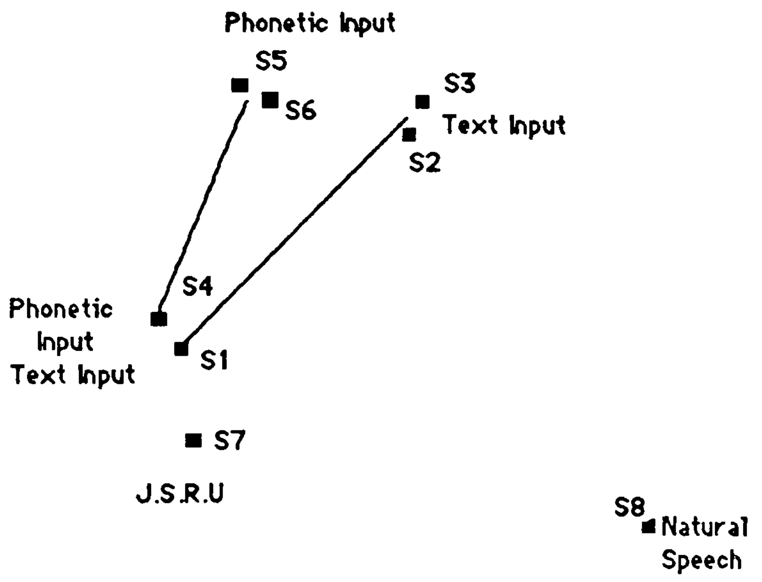


Figure (5.5)

To determine whether the modifications had any significant effect on the perceived difference between the synthetic and natural speech treatments. An analysis of variance of the mean differences data was used. The analysis is based on the mean scores shown in the last column of the proximity matrix of mean subjective scores.

Analysis of Variance of Distances from Natural Speech Treatment

Nat. Difference - ANOV.

Source of varianace	Sum of Squares	Degrees of freedom	Mean square	F - Ratio	Sig.
Distances	5.8	6	0.37	0.67	N.S.
Residual	73.4	133	0.55		
Total	79.2	139			

Table (5.5)

The ANOV shows there is no significant difference between the differences / distances of the individual synthetic treatments from the natural treatment. Hence it can be said that the modifications to S1 and S4 had no significant effect on the "perceived difference" of these two treatments when compared to the natural treatment.

Treatments S1 to S6 were all synthesized using the British Telecom system. Treatments S1 and S4 were text and phonetic input respectively. Treatments S2 &S3 are modifications of S1 text input, and treatments S5 & S6 are modifications of S4 phonetic input. From figure (5.5) it can be that the perceived differences between the modified text-input treatments S2 & S3, and the modified phonetic-input treatments are greater than the perceived difference between treatments S1 and S4 on which they are based. The modifications seem to have accentuated the difference between the text and phonetic input treatments.

The question of what attributes of speech subjects are using to differentiate between speech treatments remains. The MDS. configuration shows the relationship between the treatments based on the perceived differences between treatment pairs. From the grouping of the treatments certain intuitive conclusions can be drawn as to what subjects used to differentiate between treatments. The differences between the natural treatment and the synthetic treatments may be based on the differences of naturalness or quality of the speech. But what are subjects basing their opinion of the differences between the different types of synthetic speech. The mean results of the two SDS tests were used in an attempt to interpret the MDS. configuration.

5.2.4 Interpretation of Multidimensional Scaling Configuration

The MDS. part of the experiment yields a configuration based on the mean scores of subjects perceived differences between the between the different types of speech (treatments). The configuration is a 2 - dimensional representation of the spatial relationship between the treatments as perceived by the subjects. Thus the configuration is a simplified model of the subjects internal representation of the relationship between the treatments. The internal representation is thought to be based on a multidimensional perceptual space. The semantic differential scaling parts of the experiment give a subjective measurement of four attributes of the treatments under test. If we consider the semantic differential scales as axes in multidimensional space and the subjects mean scores as coordinates in theory the treatments could be plotted in this space . It is possible that two or more of the semantic differential scales are measuring the same fundamental attribute of speech, if this is the case the four dimensions of this theoretical space can be reduced. Principal component analysis determines the main components of variance in a set of data app. (A.2.2) and was used to see if the data from the four semantic differential scales could be reduced to fewer components i.e. fewer dimensions. The data from the SDS experiments was analysed using principal component analysis.

5.2.5 Principal Component Analysis

The mean values of the two repetitions of the SDS tests were analysized using principal component analysis the results are shown below in table (5.6).

Principal Component Analysis MDS-SDS Test 1

Treatments	Comp.1	Comp.2	Comp.3
S1	0.9979	0.0434	-0.0488
S2	0.9995	-0.0211	0.0239
S3	0.9940	-0.1089	0.0071
S4	0.9868	-0.0479	-0.1548
S5	0.9558	-0.2827	0.0802
S6	0.9956	-0.0543	-0.0769
S7	0.9878	-0.0419	0.1503
S8	0.7250	0.6881	0.0301
Component Variance	7.363	0.575	0.063
% of total variance	92.032	7.18	0.785
Total % accounted for	100		

Table (5.6)

The three components were treated as axes x, y, z of a three dimensional space and the treatment correlations were treated as the coordinates locating the treatments in this space. The distance between each treatment was calculated.

Distances between treatments based on three principal components and component correlations.

S2	S3	S4	S5	S6	S7	S8	
S1	0.97	1.62	1.40	3.53	1.02	5.04	7.05
S2		0.90	1.81	2.71	1.06	4.18	7.60
S3			1.73	1.92	1.00	3.42	8.41
S4				3.34	0.79	4.80	8.03
S5					2.80	1.57	9.99
S6						4.30	7.97
S7							11.44

The matrix of calculated distances between treatments was used as input data to the multidimensional scaling analysis program. The program attempts to produce a two dimensional configuration where the distances between the points on the configuration approximates the input distances.

Distances fitted to mean of S.D.S.1&2 principal component distances using multidimensional scaling analysis program.

	S2	S3	S4	S5	S6	S7	S8
S1	0.93	1.65	1.40	3.57	1.00	5.11	7.14
S2		0.90	1.68	2.84	0.92	4.40	8.06
S3			1.63	1.95	0.95	3.50	8.62
S4				3.02	0.78	4.43	7.20
S5					2.68	1.56	10.22
S6						4.19	7.69
S7							11.56

The coordinates of the treatments in the two dimensional configuration are;

S1 (-0.49, -0.99)	S5 (-1.16, 2.51)
S2 (-1.14, -0.33)	S6 (-0.27, -0.01)
S3 (-1.02, 0.57)	S7 (-1.21, 4.07)
S4 (0.51, -0.01)	S8 (4.77, -5.82)

A plot of the configuration is shown below, figure (5.6)

MDS Configuration of Principal Component Data Mean of SDS1&2 Test 1

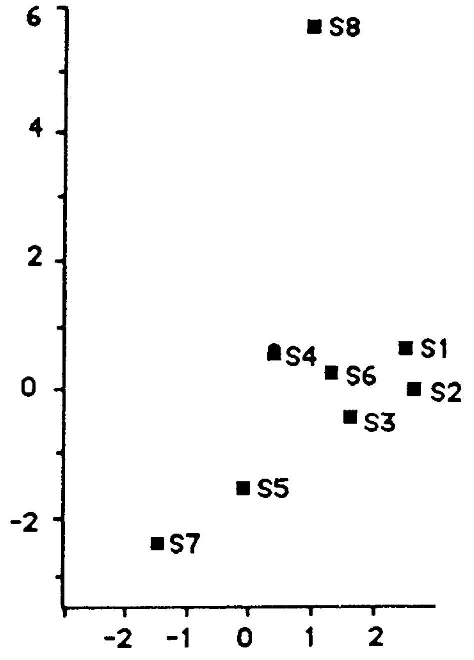


Figure (5.6)

The configurations from the MDS data and the principal component data can be compared by standardizing the configurations and plotting them on the same axes, figure (5.7).

Comparison of MDS and Principal Component Configurations Test 1.

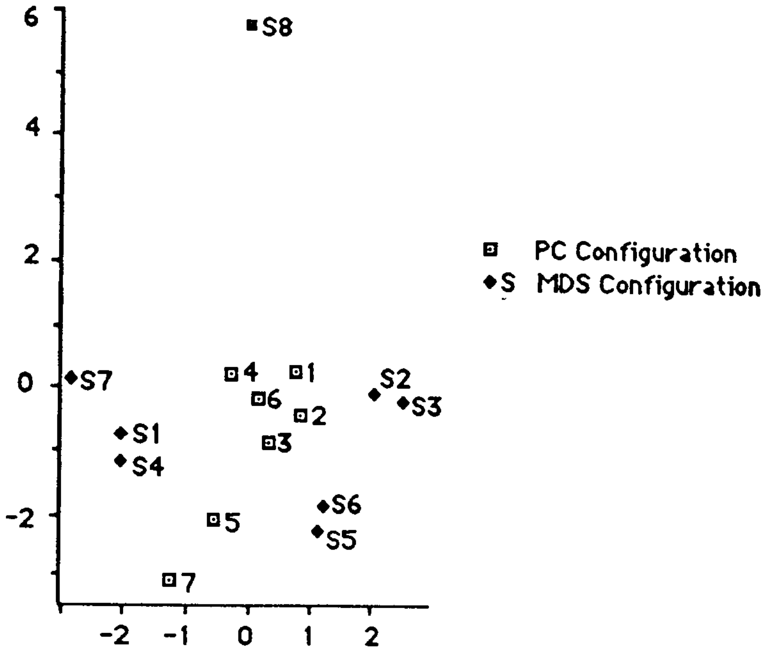


Figure (5.7)

An attempt was made to interpret the MDS configuration in terms of the attributes measured in the SDS tests. The rank order of the treatments based on each Semantic differential scale scores were used. For each attribute i.e. naturalness a line was plotted on the configuration which passed through the centre of gravity of the configuration. The line was rotated so that the position of each treatment (measured at 90 degrees to the line) matched the rank order of treatments rated on the naturalness semantic differential scale. An interpretation of the MDS configuration in terms of the four attributes measured in the SDS tests is shown in figure (5.8) below.

Interpretation of MDS Configuration

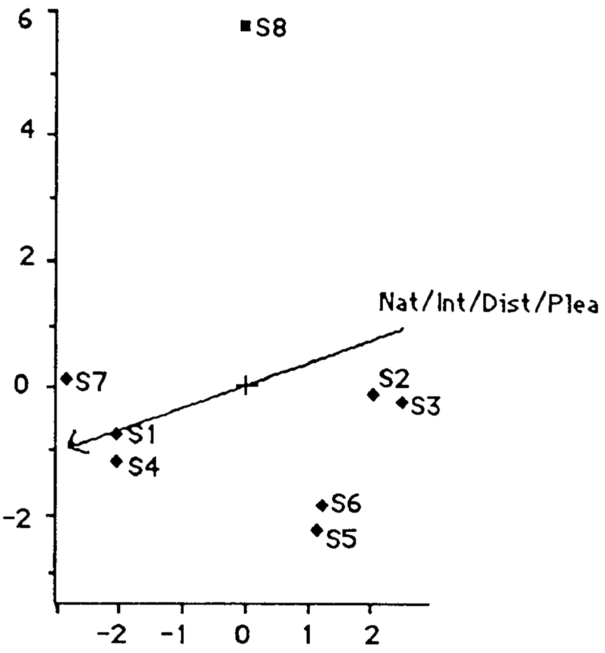


Figure (5.8)

The synthetic speech treatments were perceived to be very different from the natural speech. The scores for specific attributes of the synthetic treatments were in general quite low. Hence the natural speech was not included when aligning the results from the SDS tests with the configuration. Another point to note is that in general the synthetic speech treatment SDS scores were not significantly different which made aligning the attribute scales with the configuration less accurate. From figure (5.8) it can be seen that all of the SDS (attribute) scales are all in one direction. Corresponding with the modified and unmodified synthetic treatments showing. This indicates that the modified synthetic treatments were degraded compared to the unmodified treatments.

5.2.6 Comparison of Subjects M.D.S. configurations.

A standard configuration was used to compare the subject's M.D.S. configurations. The standard configuration is based on measuring the perceived differences between the three types of speech rather than the differences between all treatment pairs. This was done by calculating the distances between the centre of gravity for the three unmodified synthesized speech treatments, the centre of gravity of the modified synthesized speech treatments, and the natural treatment. The scale used by subjects to score their opinion of the perceived difference was ten units long. Hence the maximum perceived difference between two treatments was limited to ten units. It therefore follows that the maximum distance between the two centres of gravity and the natural treatment could not be more than ten units. This extreme case can be represented as a triangle fig(5.9a) where the three distances a, b, and c equal ten units. The centre of gravity of the configuration C.g. is set at the origin of the cartesian coordinates. The distance r of the treatment centres of gravity is 5.774 units. Syn1 is the centre of gravity of the three unmodified synthetic treatments. Syn2 is the centre of gravity of the four modified synthetic treatments.

The standard configuration can be represented by figures (5.9 a-b).

MDS Standard Configuration

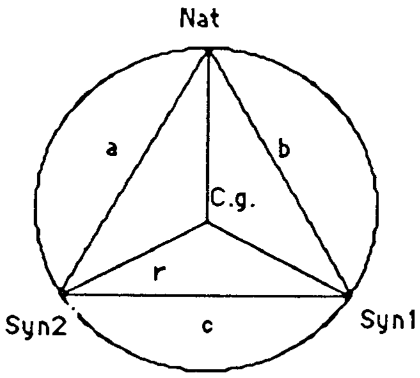


Figure (5.9a)

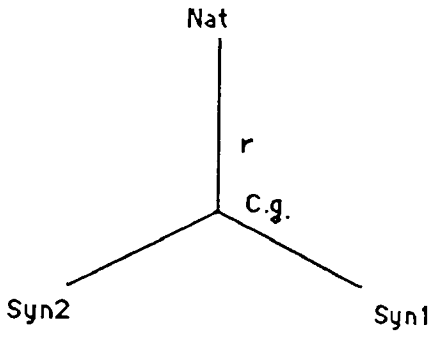


Figure (5.9b)

The standard configuration provides a simple frame of reference for comparing subject's M.D.S. configurations. The subject's M.D.S. configurations were transformed into polar coordinates and rotated so that the natural treatment was at 90 degrees. The centres of gravity of the modified and unmodified synthetic speech treatments were calculated. If necessary configurations were flipped about the y - axis to aline them with the standard configuration. An example of the difference between two subjects standardized configurations is shown in figure (5.10).

Standard Configurations of Subjects 1 & 2

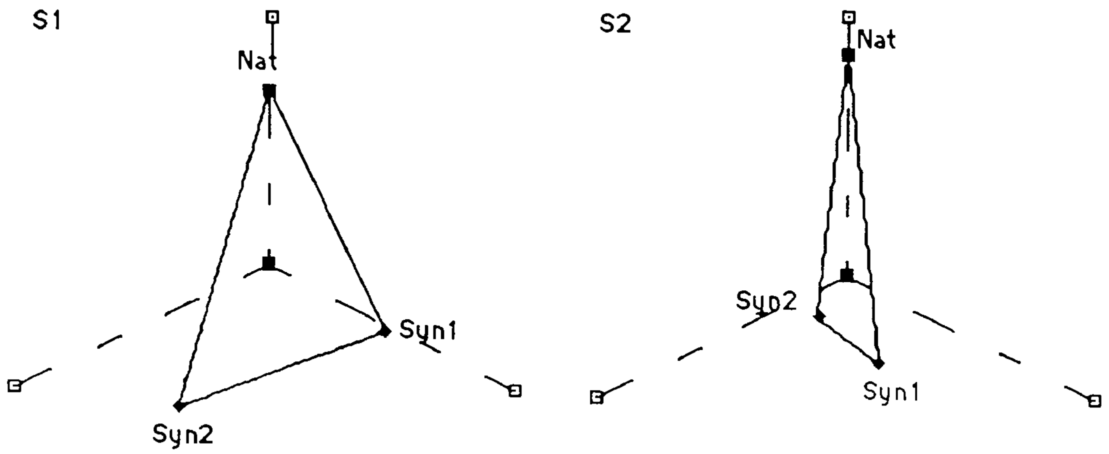


Figure (5.10)

For all subjects the distances between the three types of speech were calculated and are shown below. Distances between the three types of speech from standardized configurations are shown below in table (5.7), i.e. column 'Nat - Syn1' is the distance between the natural and unmodified synthetic speech treatments.

	Nat - Syn1	Nat - Syn2	Syn1 - Syn2
	6.05	7.57	4.51
	7.27	6.14	1.63
	6.81	6.82	3.45
	6.32	6.91	1.96
	7.27	6.4	3.29
	6.88	6.86	3.83
	6.56	7.22	4.39
	7.11	7.25	5.29
	6.96	7.07	4.89
	6.7	6.68	2.27
	6.98	6.75	3.74
	8.29	5.06	7.55
	7.13	7.19	5.42
	6.79	6.96	4.0
	6.6	7.37	5.03
	6.74	7.01	4.06
	9.0	4.8	4.29
	6.81	6.98	4.08
Mean Distance	7.02	6.72	4.09
Variance	0.46	0.52	1.90

Table (5.7)

An analysis of variance was used to determine whether the means of the differences between the three types of speech were significantly different. The distance means were compared and the F ratios calculated.

	Nat - Syn2	Syn1 - Syn2
Nat - Syn1	1.62 - NS	68.08 - ***
Nat - Syn2		49.41 - ***

It can be seen that the mean perceived difference between the two types of synthetic speech is significantly different from the perceived differences between the natural speech and the two types of synthetic speech. The homogeneity of the variances of the three mean perceived differences was tested using Bartlett's test, app. (A.2.5). The test value 'q' of the three variances was calculated (q=8.599) compared to the chi-square 95% critical value of 5.99. The test value is higher, hence the variances are not homogeneous. The variances were all compared to each other, the q values are shown below.

	Nat - Syn2	Syn1 - Syn2
Nat - Syn1	-9.0925	5.76
Nat - Syn2		6.53

95% critical value from chi-squared tables 3.84.

The variances of the perceived differences of the two types of synthetic speech from the natural speech are homogeneous. Where as the variance of the perceived difference between the two types of synthetic speech is significantly different from the variances of the other two perceived differences. The variance between Syn1 and Syn2 is significantly greater than the other variances. This means that there was a greater range of subject scores, perceived differences, between the synthetic treatments.

One of the objectives of this MDS and SDS experiment was to determine whether simple modifications to the phoneme parameters of the British Telecom synthesis system would improve the quality of the synthetic speech. The analysis of variance of the mean distances of the synthetic treatments from the natural treatment sec. (5.2.3). Showed that there was no significance between the modified and unmodified synthetic treatments. The least significant difference diagrams sec. (5.2.1) show that S4, the unmodified phonetic input treatment was ranked highest on all four semantic differential scales. Hence in later experiments the British Telecom synthesis system was used with it's default phonetic parameters.

5.3 Results of Second Multidimensional (M.D.S.) and Semantic Differential Scaling (S.D.S.) Experiment

5.3.0 General

The purpose of this experiment was to investigate the perceived differences between synthetic speech and TFW modulated speech. and to compare specific attributes of the two types of speech.

5.3.1 Speech Material and Semantic Differential Scales

The experimental procedure described in sec. (4.4.2.1) was used. The speech treatments were;

S1 - PDP11 Text input,no modifications

S2 - PDP11 Text input, Fixed pitch contour

S3 - TFW,Modulation: Sine; Period; 20 ms; Amplitude 2 KHz

S4 - TFW,Modulation: Sine; Period; 150 ms; Amplitude 1.5 KHz

S5 - TFW,Modulation: Sine; Period; 200 ms; Amplitude 2 KHz

S6 - MACTALK, Phonetic corrections

S7 - Natural speech

The sentences used in all of the treatments were;

"He carried a bag of tennis balls."

Closely followed by;

"It did not seem like summer."

These two sentences were used in both the MDS. and SDS tests because it was the difference between the types of speech, we wanted the subjects to assess, and not the effect of sentence / voice combinations.

Five semantic differential scales were used;

Intelligibility	- Is it easy or hard to understand separate words in the speech
Naturalness	- Does the speech sound natural or unnatural
Distinctness	- does the speech sound clear or slurred
Pleasantness	- Does the speech sound pleasant or unpleasant
Listening Effort	- (absence of) Effort required to understand the meaning of the speech , no effort to considerable effort.

5.3.2 Results of Semantic Differential Scaling Tests.

Table (5.8) shows the mean subjective scores for the two S.D.S. tests. Plus the mean of the two repetitions. Scores were measured from the semantic differential scales eg, Unnatural = 0, Natural = 10

Treatment Mean Subjective Scores for Semantic Differential Scales MDS-SDS Test 2

		S1	S2	S3	S4	S5	S6	S7
Intelligibility	Rep.1	5.32	4.96	4.85	3.18	1.78	3.54	9.29
	Rep.2	5.17	4.80	4.64	3.99	3.04	5.11	9.49
	Mean	5.25	4.88	4.75	3.59	2.41	4.33	9.39
Naturalness	Rep.1	1.68	1.50	3.77	2.42	1.28	2.12	9.18
	Rep.2	2.64	1.58	3.09	4.11	2.84	2.33	9.72
	Mean	2.16	1.54	3.43	3.27	2.06	2.23	9.45
Distinctness	Rep.1	2.87	4.91	3.54	1.46	1.34	2.51	9.26
	Rep.2	6.47	5.92	3.29	2.53	1.71	5.68	9.29
	Mean	4.67	5.42	3.42	2.00	1.52	4.10	9.28
Listening Effort	Rep.1	3.89	4.92	3.99	2.91	1.94	2.88	9.04
	Rep.2	4.57	4.67	4.52	3.64	3.01	4.63	9.28
	Mean	4.23	4.80	4.26	3.28	2.48	3.76	9.16
Pleasantness	Rep.1	1.69	3.13	2.48	2.02	1.67	2.28	9.03
	Rep.2	2.38	1.47	3.83	3.64	3.39	2.03	9.03
	Mean	2.04	2.30	3.16	2.83	2.53	2.16	9.30

Table (5.8)

The the mean scores of the two repetitions for each of the seven speech treatments were plotted as a column graph fig.(5.11). The graph provides a simple profile of each treatment based on the five attributes measured on the semantic differential scales.

Treatment Profiles MDS-SDS Test 2

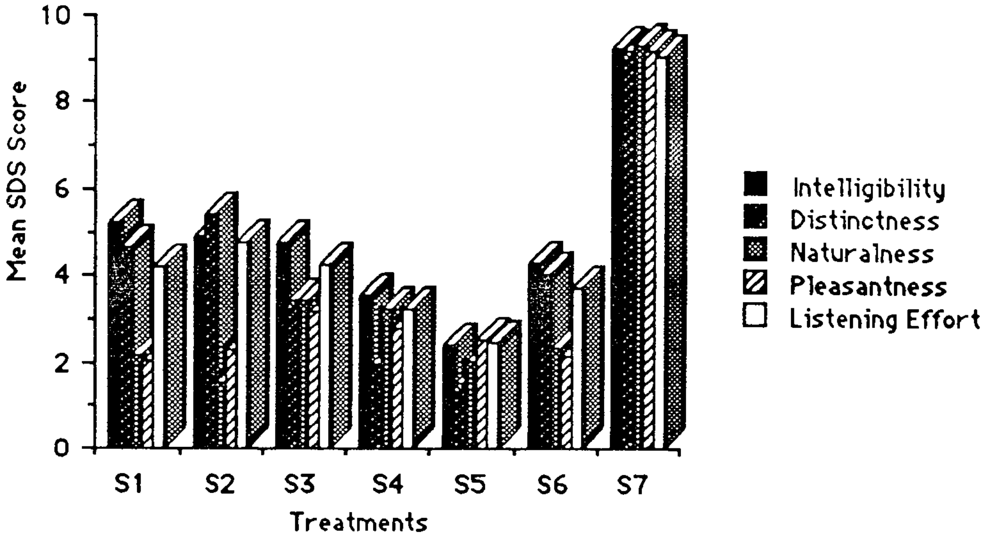


Figure (5.11)

5.3.3 Analysis of variance of the treatment means.

A one - way analysis of variance was used to compare the treatments mean scores for the five semantic differential scales. The significance of the difference between the treatment means was calculated. Also the least significant difference between the treatment means was calculated to determine the grouping of the treatments. The analysis is based on treatment mean scores calculated from combined data from both repetitions. The results are shown in tables (5.9 a-e) below, *** indicates that the treatment means are significantly different and NS indicates they are not significantly different.

Analysis of Variance of Semantic Differential Scales MDS-SDS Test 2

ANOV - Intelligibility - SDS

Source or Variance	Sum of Squares	Degrees of Freedom	Mean Squared	F - ratio	Sig.
Treatments	97.0	5	19.4	5.22	***
Residual	379.3	102	3.72		
Total	476.1	107		L.S.D.	1.26

Homogeneous Groupings	S1	S2	S3	S6	S4	S5
	5.25	4.88	4.74	4.33	3.64	2.41
	▬	▬	▬	▬		
		▬	▬	▬	▬	
					▬	▬

Table (5.9a)

ANOV - Naturalness - SDS

Source or Variance	Sum of Squares	Degrees of Freedom	Mean Squared	F - ratio	Sig.
Treatments	72.35	5	14.47	3.84	***
Residual	384.15	102	3.77		
Total	456.5	107		L.S.D.	1.27

Homogeneous Groupings	S3	S4	S6	S1	S5	S2
	3.43	3.27	2.23	2.18	2.06	0.98
	▬	▬	▬	▬		
		▬	▬	▬	▬	
			▬	▬	▬	▬

Table (5.9b)

ANOV - Distinctness - SDS

Source or Variance	Sum of Squares	Degrees of Freedom	Mean Squared	F - ratio	Sig.
Treatments	95.5	5	19.1	1.81	N.S.
Residual	1076.74	102	10.56		
Total	1172.24	107		L.S.D.	2.12

Homogeneous Groupings	S6	S2	S3	S1	S4	S5
	3.96	3.71	3.42	2.27	1.99	1.55
	▬	▬	▬	▬	▬	
			▬	▬	▬	▬

Table (5.9c)

ANOV -Pleasantness - SDS

Source or Variance	Sum of Squares	Degrees of Freedom	Mean Squared	F - ratio	Sig.
Treatments	14.58	5	2.92	1.15	N.S.
Residual	259.28	102	2.54		
Total	288.49	107		L.S.D.	1.04

Homogeneous	S3	S4	S5	S2	S6	S1
Groupings	3.16	2.83	2.5	2.3	2.16	2.03
	▬	▬	▬	▬	▬	
		▬	▬	▬	▬	▬

Table (5.9d)

ANOV - Listening Effort - SDS

Source or Variance	Sum of Squares	Degrees of Freedom	Mean Squared	F - ratio	Sig.
Treatments	61.1	5	12.22	4.01	***
Residual	311.17	102	3.05		
Total	372.35	107		L.S.D.	1.14

Homogeneous	S2	S3	S1	S6	S4	S5
Groupings	4.79	4.26	4.23	3.7	3.28	2.48
	▬	▬	▬	▬		
		▬	▬	▬	▬	
					▬	▬

Table (5.9e)

5.3.4 Determination of Perceived Differences Between TFW Modulated Synthetic and Natural Speech

Subjects scores of the perceived difference (proximity) between two treatments for all of the treatment pairs were used to calculate the mean proximities between the treatments. The resultant mean proximity matrix is shown below. This was used as the input to the multidimensional scaling analysis program app. (A.3.1).

Mean Proximity Data Matrix

	S2	S3	S4	S5	S6	S7
S1	1.80	6.66	7.10	5.9	1.68	6.67
S2		7.48	7.64	7.61	2.66	7.93
S3			3.11	2.42	7.47	4.19
S4				0.46	7.50	4.11
S5					7.94	3.94
S6						8.77

Distances Fitted to Mean Proximity Data

	S2	S3	S4	S5	S6	S7
S1	1.96	6.37	6.68	6.38	1.52	6.84
S2		7.57	7.33	7.14	2.17	8.5
S3			2.23	1.79	7.83	2.69
S4				0.48	8.20	4.91
S5					7.89	4.47
S6						8.05

"Goodness of fit" Stress = 0.0974 The stress measures how well the fitted distance matrix matches the mean proximity matrix. According to theory a stress value below 0.2 is acceptable, 0.0974 is a very good fit. The coordinates of the treatments from which the fitted distances were calculated are;

S1	(-3.24, 0.29)
S2	(-3.89, 2.13)
S3	(3.00, -0.98)
S4	(3.38, 1.21)
S5	(3.12, 0.81)
S6	(-4.75, 0.14)
S7	(2.38, -3.60)

Plot of Configuration.

MDS Configuration of Subjects Mean Perceived Differences Test 2

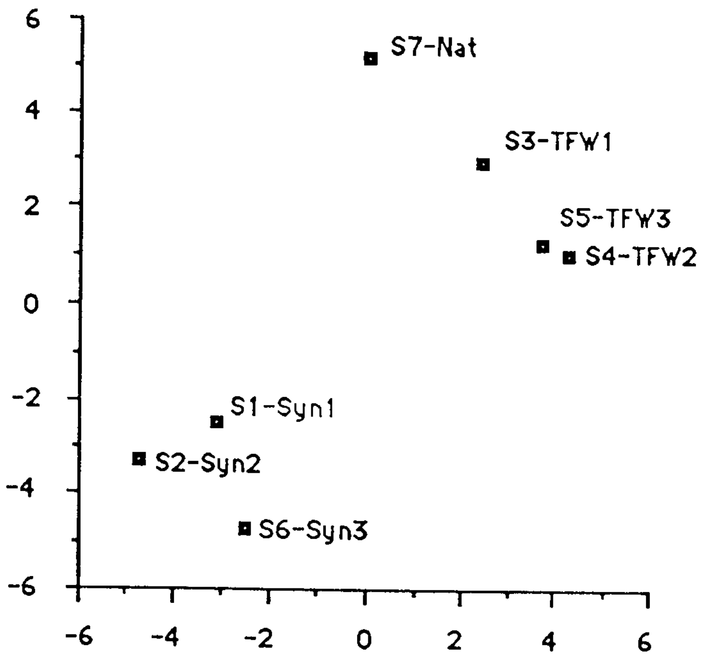


Figure (5.12)

Figure (5.12) shows the 2 - dimensional spatial relationship of the treatments based on the subjects mean scores of the perceived differences (proximities) between the treatments. The configuration is shown with the natural treatment S7 at 90 degrees to the configuration's centre of gravity (0, 0). The orientation of the configuration is not important and the units of the axes are arbitrary, as long as the relative distances between treatments are preserved. So the configuration can be rotated and flipped about the axes etc. Rotating the configuration so that S7 natural treatment is at 90 degrees to the centre of gravity enables us to use the line from the centre of gravity to S7 as a reference when comparing configurations. It is desirable to be able to compare the individual subjects configurations. One method is to use a standard configuration.

5.3.5 Comparison of Subjects MDS. Configurations Test 2

Subjects 3&4 Standard Configurations Test 2

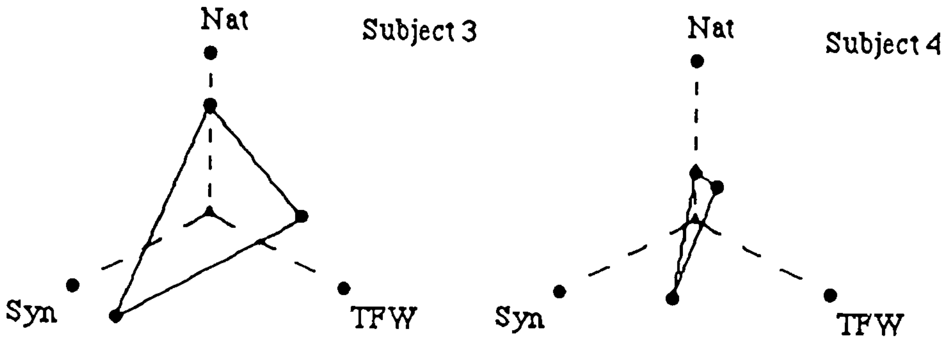


Figure (5.13)

Figure (5.13) shows the simplified MDS configurations for subjects 3 and 4. It can be seen that subject 3's perceived difference between the three types of speech was greater than that of subject four. Further information was gained about the subjects perceived differences between the three types of treatments, by calculating the mean value of each difference and comparing them using an analysis of variance. The distances from subjects standard configurations are shown in table (5.10).

Data from Subjects Standard Configurations MDS-SDS Test 2

Subjects	Nat - Syn	Nat - TFW	Syn - TFW
1	9.47	5.76	9.44
3	8.45	5.27	7.81
4	4.73	1.00	4.40
5	7.47	3.41	7.42
6	6.01	3.10	6.74
7	10.01	2.21	8.82
8	8.82	4.32	6.16
9	7.41	3.91	6.48
10	6.39	3.16	6.11
Mean	7.68	3.57	7.04

Table (5.10)

An analysis of variance was carried out on the mean distances data and the least significant difference calculated, table (5.11).

MDS Standard Configurations Analysis of Variance MDS-SDS Test 2

ANOV -STD. Configurations

Source or Variance	Sum of Squares	Degrees of Freedom	Mean Squared	F - ratio	Sig.
Distances	88.03	2	44.02	16.86	***
Residual	62.70	24	2.61	L.S.D.	1.62

Table (5.11)

The F - ratio for two variables with degrees of freedom 2 & 24 are significantly different at the 95 percent level of significance if 'F' exceeds 3.4. Hence it can be seen that the three distances are significantly different. To clarify the differences the least significance difference was calculated. The least significance difference was 1.62. This figure was compared with the differences between the three mean distances. The difference between Nat - Syn and Syn - TFW are not significantly different whereas Nat - Syn and Syn - TFW are both significantly different from Nat - TFW. Another factor to be considered is the variance of the distances. The homogeneity of the variances can be tested using Bartlett's test. The variances of the distance means were.

1	Nat - Syn	Var. - 3.23
2	Nat - Tfw	Var. - 2.15
3	Syn - TFW	Var. - 2.33

The result of Bartlett's test was 0.411 this was compared with the chi-squared critical value of 5.99. This showed that the variances were homogeneous, i.e. the variances are not statistically different.

5.3.6 Interpretation of Multidimensional Scaling Configuration

The aim of the MDS. and SDS. experiments was to try to determine which attributes of speech subjects used in determining the perceived difference between samples of speech (treatments). The MDS. part of the experiment yields a configuration based on the mean scores of subjects perceived differences between the different types of speech. The configuration is a 2 - dimensional representation of the spatial relationship between the treatments as perceived by the subjects. Thus the configuration is a simplified model of the subjects internal representation of the relationship between the treatments. The internal representation is thought to be based on a multidimensional perceptual space.

The SDS. parts of the experiment give a subjective measurement of five attributes of the treatments under test. If we consider the semantic differential scales as axes in multidimensional space and the subjects mean scores as coordinates in theory the treatments could be plotted in this space . It is possible that two or more of the semantic differential scales are measuring the same fundamental attribute of speech, if this is the case the five dimensions of this theoretical space can be reduced. Principal component analysis determines the main components of variance in a set of data app. (A.2.2) and was used to see if the data from the five semantic differential scales could be reduced to fewer components i.e. fewer dimensions. The data from the two SDS. experiments and the data of the mean of the experiments were analysed using principal component analysis.

5.3.7 Principal Component Analysis of Semantic Differential Scales

The mean values of the two repetitions of the SDS tests were analysized using principal component analysis the results are shown below in table (5.12).

Principal Component Analysis MDS-SDS Test 2

Treatments	Comp.1	Comp.2	Comp.3	Comp.4
S1	0.9470	0.2524	0.1956	-0.0352
S2	0.9969	-0.0389	0.0476	-0.0486
S3	0.6104	0.7808	-0.1164	0.0644
S4	-0.1899	0.9242	-0.3106	0.1153
S5	0.6970	-0.6691	0.0049	0.2578
S6	0.9608	0.1999	0.1790	-0.0695
S7	-0.5495	0.4133	0.7207	0.0884

Component Variance	4.010	2.188	0.702	0.100
Percentage of total variance	57.29	31.25	10.03	1.43
Total percentage accounted for 100 %				

Table (5.12)

The first three principal components accounted for most or all of the variance in the data. If we consider the three components as axes x,y, and z and the correlation of each treatment with the three components are considered as the coordinates of the treatments in 3-dimensional space, the relative distances between treatments can be calculated. It would be useful to be able to reduce this 3 - dimensional semantic differential scaling configuration to 2 - dimensions. This can be done by either looking at two components at a time or by a 2 - dimensional configuration which models the 3 - dimensional configuration. The multidimensional scaling (MDS.) analysis program is designed to model multidimensional space based on distance input data. This was used to reduce the 3 - dimensional configuration to two dimensions. The distance between each treatment in the three dimensional configuration was calculated and used as the input data for the MDS. analysis program.

5.3.7.1 MDS. Analysis of of the Mean of First & Second Semantic Differential Scaling Tests.

Calculated distances between treatments using the first three principal components as x, y,z axes and treatment correlations as coordinates. Calculated distance input matrix.

	S2	S3	S4	S5	S6	S7
S1	1.56	4.59	12.44	3.14	0.22	15.86
S2		4.20	12.40	3.03	1.36	16.87
S3			8.24	1.49	4.58	14.30
S4				9.41	12.51	10.92
S5					3.16	14.37
S6						16.05

Fitted distances

	S2	S3	S4	S5	S6	S7
S1	1.57	4.56	12.45	3.13	0.22	15.90
S2		4.2	12.39	3.02	1.37	16.90
S3			8.21	1.46	4.56	14.21
S4				9.42	12.52	10.89
S5					3.15	14.38
S6						16.09

Stress after 50 iterations 0.0065

Coordinates of treatments in 2-dimensional configuration from which fitted distances were calculated.

S1 (-4.54, 0.90)	S5 (-2.02, -0.96)
S2 (-5.02, -0.60)	S6 (-4.68, 0.73)
S3 (-1.1, -2.1)	S7 (10.48, 6.15)
S4 (6.86, -4.12)	

The configurations from the multidimensional scaling data and the principal component data can be compared by standardizing the configurations and plotting them on the same axes, figure (5.14). Standardizing the configurations in this case means rotating the configurations so that the natural speech treatment lies on the Y - axis, and flipping the configuration about the Y- axis so that the synthetic speech treatments of both configurations are on the same side.

Comparison of the MDS and SDS Based Configurations Test 2

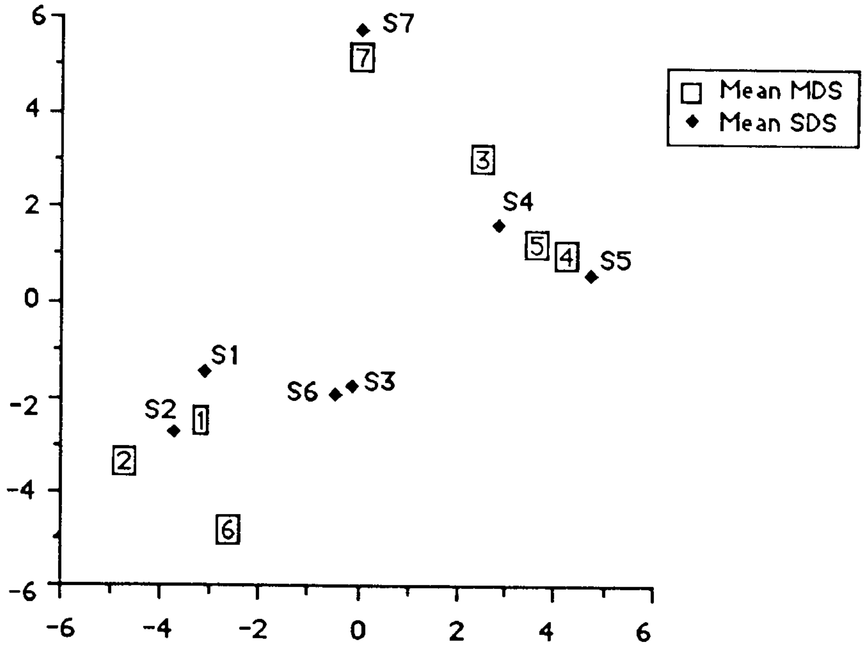


Figure (5.14)

The two configurations match up surprisingly well considering they are based on different types of data. The multidimensional scaling MDS configuration is based on the perceived differences between treatments. Whereas the semantic differential scaling SDS configuration is based on subjective measurements of specific attributes of the treatments. The similarity of the configurations suggests that subjects were basing their opinions of the perceived difference between treatments on the attributes measured in the SDS tests, i.e. Naturalness, pleasantness, intelligibility, distinctness and listening effort. It should therefore be possible to interpret the MDS configuration in terms of the attributes measured in the SDS tests. The rank order of the treatments based on each Semantic differential scale scores were used. For each attribute i.e. naturalness a line was plotted on the configuration which passed through the centre of gravity of the configuration. The line was rotated so that the position of each treatment (measured at 90 degrees to the line) matched the rank order of treatments rated on the naturalness semantic differential scale. An interpretation of the MDS configuration in terms of the five attributes measured in the SDS tests is shown in figure (5.15) below.

Interpretation of MDS Configuration in Terms of Speech Attributes

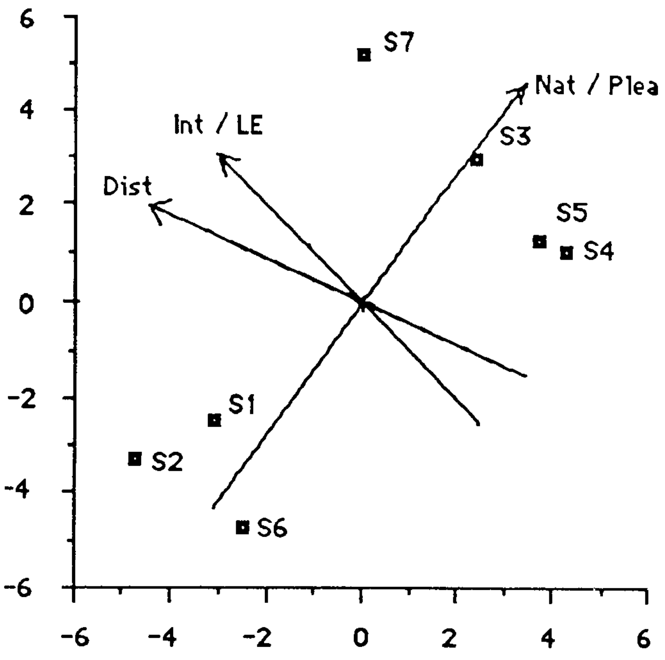


Figure (5.15)

5.4 Listening Effort Test to Determine TFW Variable Parameter

5.4.0 General

Having determined which TFW modulation waveform was to be used, the next step was to determine whether the modulation period or amplitude was to be used as the variable parameter sec (4.1). A listening effort test was carried out using different combinations of period and amplitude settings. The eleven treatments used in the test are shown below:-

Treatments

R1 - Natural speech

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

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

5.4.1 Results and Analysis

The mean treatment scores based on 11 subjects over the five listening levels were calculated.

TFW Parameter Listening Effort Test Mean Scores Repetition 1

R1	T1	T5	T2	T4	S3	T3	S1	S2	T6	T7
3.71	2.93	2.67	2.20	1.69	1.65	1.33	1.04	0.98	0.80	0.22

TFW Parameter Listening Effort Test Mean Scores Repetition 2

R1	T1	T5	T2	T4	S3	S2	T3	S1	T6	T7
3.84	3.65	3.20	2.64	2.13	1.84	1.51	1.42	1.38	1.05	0.27

The treatments were plotted as column graphs, figs. (5.16a-b).

Column Graphs of Treatments

First Repetition

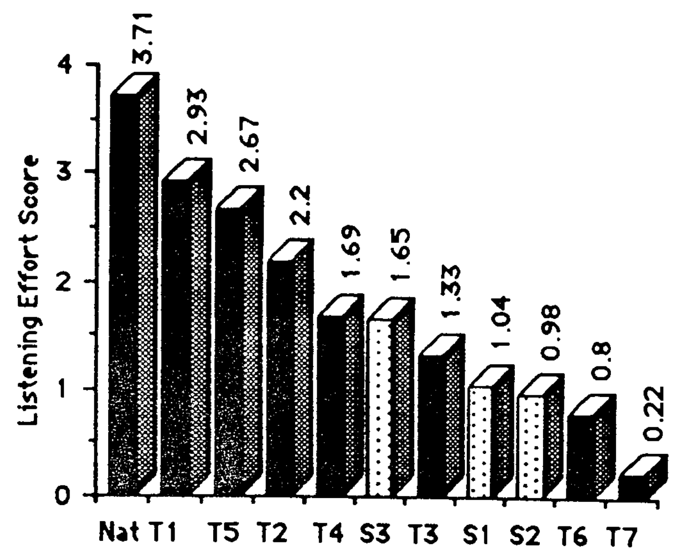


Figure (5.16a)

Second Repetition

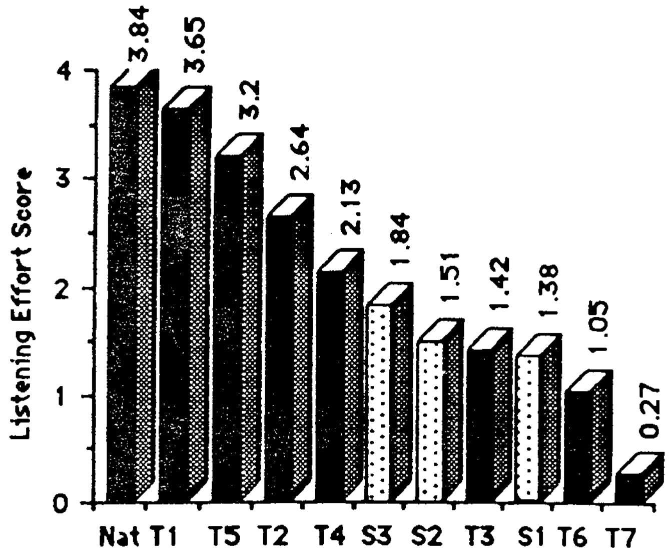


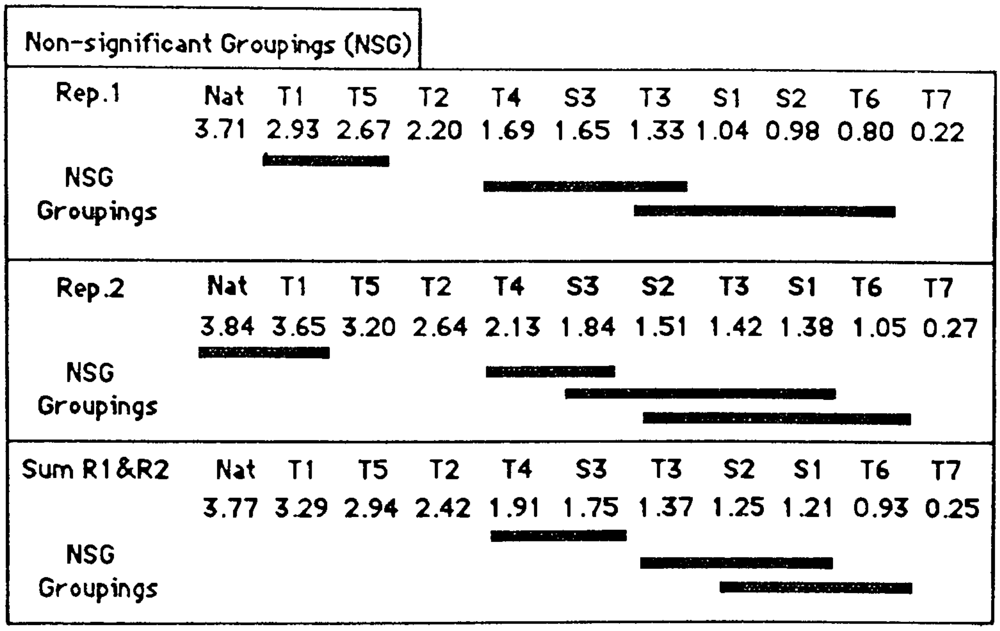
Figure (5.16b)

5.4.1.1 Groupings of Not Significantly Different Treatments

It can be seen from the column graphs figs. (5.16a-b), that except for treatment S2 the rank order of the treatments remains the same for both repetitions of the comparison listening effort test. To determine whether the change of position of S2 is significant.

Plus determine groupings of treatments that are not significantly different. The F-ratio between all combinations of treatment pairs were calculated. Treatment means that were not significantly different i.e. the F-ratio is below the critical value, are indicated by drawing a line under that group Figure (5.17) shows the grouping of treatments that are not significantly different.

Non Significant Groupings of Treatments



Non-significant Groupings (NSG)

Rep.1	Nat	T1	T5	T2	T4	S3	T3	S1	S2	T6	T7
	3.71	2.93	2.67	2.20	1.69	1.65	1.33	1.04	0.98	0.80	0.22
NSG Groupings		▬	▬								
					▬	▬	▬				
							▬	▬	▬	▬	

Rep.2	Nat	T1	T5	T2	T4	S3	S2	T3	S1	T6	T7
	3.84	3.65	3.20	2.64	2.13	1.84	1.51	1.42	1.38	1.05	0.27
NSG Groupings	▬	▬									
					▬	▬					
						▬	▬	▬	▬		
							▬	▬	▬	▬	

Sum R1&R2	Nat	T1	T5	T2	T4	S3	T3	S2	S1	T6	T7
	3.77	3.29	2.94	2.42	1.91	1.75	1.37	1.25	1.21	0.93	0.25
NSG Groupings					▬	▬					
							▬	▬	▬		
								▬	▬	▬	

Figure (5.17)

5.4.1.2 Complete Graeco Latin Square x Layers Analysis of Variance

First Repetition

Factor	d.f.	SS	MS	F-ratio	Sig
Subjects	10	48.67	4.87	3.244	**
Sequence	10	11.00	1.10	0.733	NS
Treatments	10	594.96	59.50	39.656	***
Lists	10	12.45	1.25	0.830	NS
Error(1)	80	120.02	1.50	3.949	***
Levels	4	3.27	0.82	2.149	NS

Interactions

Subjects x Listening Level	40	18.19	0.45	1.197	NS
Sequences x Listening Levels	40	19.68	0.49	1.295	NS
Treatments x Listening Levels	40	12.44	0.31	0.819	NS
Lists x Listening Levels	40	16.04	0.40	1.056	NS
Error(2)	320	121.58	0.38		
Total	604	978.31			

Second Repetition

Factor	d.f.	SS	MS	F-ratio	Sig
Subjects	10	41.36	4.14	3.111	**
Sequence	10	13.94	1.39	1.049	NS
Treatments	10	701.72	70.17	52.794	***
Lists	10	27.36	2.74	2.058	*
Error(1)	80	106.33	1.33	2.813	***
Levels	4	4.78	1.20	2.531	*

Interactions

Subjects x Listening Level	40	22.56	0.56	1.194	NS
Sequences x Listening Levels	40	24.34	0.61	1.288	NS
Treatments x Listening Levels	40	14.38	0.36	0.761	NS
Lists x Listening Levels	40	16.74	0.42	0.886	NS
Error(2)	320	151.19	0.47		
Total	604	1124.70			

5.4.1.3 Multiple Regression Analysis

The mean treatment score of the 11 subjects over the five listening levels was calculated.

First Repetition

R1	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
3.7	1.04	0.98	1.65	2.92	2.2	1.32	1.69	2.67	0.8	0.22

Second Repetition

R1	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
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 app.(A.2.3). The model relates the listening effort score Y to the time frequency warping modulation settings period and amplitude.

Mathematical Model

$$Y = 4.29- (5.5*10**-3(\text{period})) - (9.24*10**-4(\text{amplitude}))$$

Relationship Between TFW Parameters Listening Effort Score

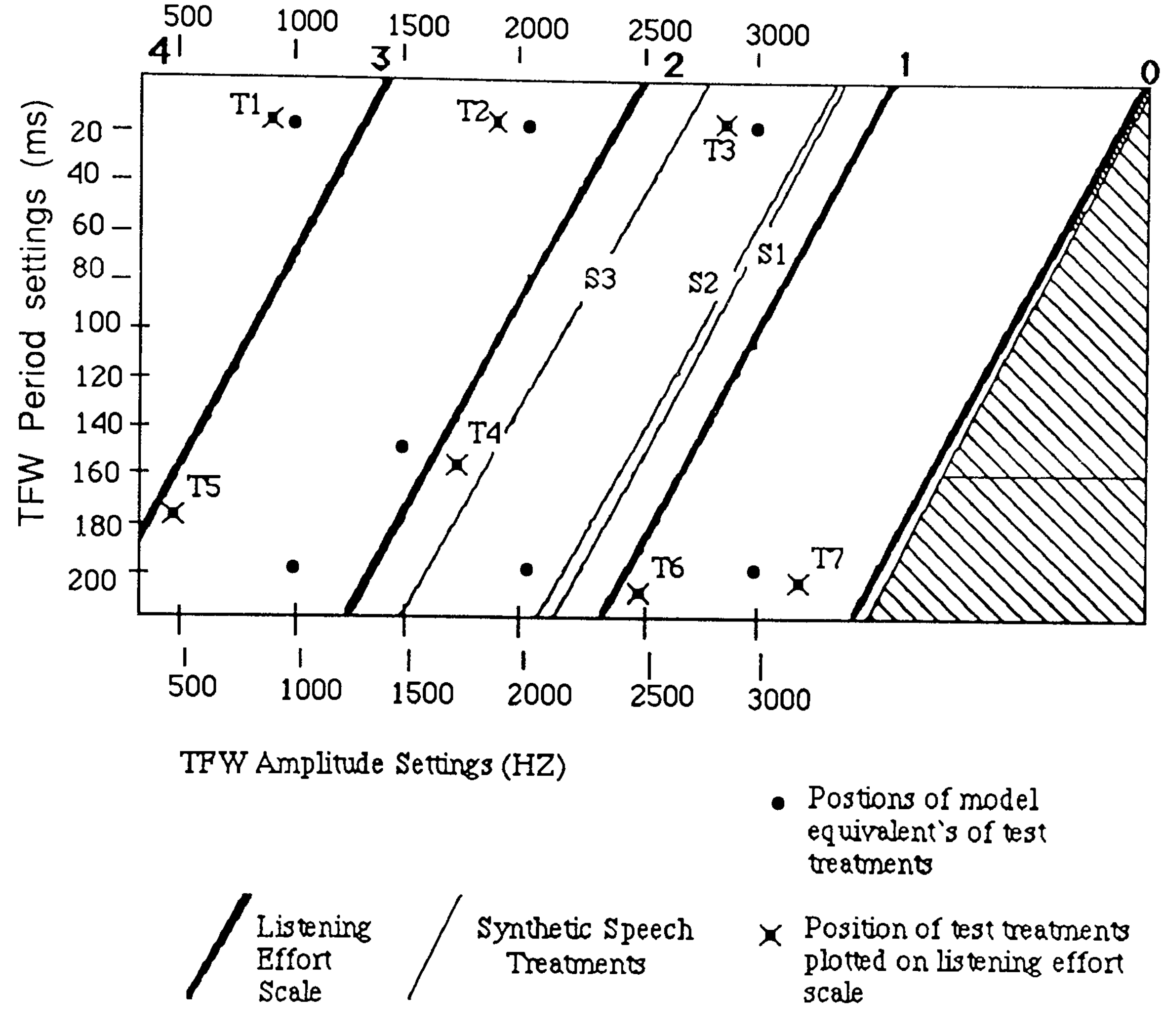


Figure (5.18)

Figure (5.18) was plotted using equation 1 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 values of Y (LE. scores) on the grid. The listening effort scale (0 - 4), is represented as thick diagonal lines. The synthetic speech treatment scores were plotted in a similar manner and are represented by the thin diaganal lines. The TFW period and amplitude settings used in the listening effort test were plotted on the grid as black dots. The subjective mean scores for the TFW modulated speech were plotted as a spiked circle on the listening effort scale. The points were plotted on a line which is 90 degrees to the listening effort scale, which passes through the model equivalent period and amplitude settings. From figure (5.18) it can be seen that for a fixed period the range of listening effort scores obtained by varying the amplitude setting is greater than the range obtained by fixing the amplitude and varying the period. Hence it was reasonable to adopt the amplitude at a fixed period as the variable parameter for TFW modulation, the period setting selected was 150 ms. From the model above the highest LE. score is 3.1 and the lowest is 0.6, this range can be increased by reducing the amplitude. Intuitively reducing the amplitude would yield higher LE. scores, because a zero amplitude setting would reproduce the original speech.

5.5 Comparison of TFW Modulated and Synthetic Speech in Terms of Listening Effort

5.5.0 General

A series of comparison listening effort tests were carried out using several examples of synthesized and TFW modulated speech. The aims of these tests were:-

i - Determine whether subjects could rate the TFW modulated speech consistently

ii - Make some measurements of synthetic speech in terms of TFW modulation equivalent settings TFWe.

iii - Determine whether there is any significant difference in the subjective variance of sample means for the TFW modulated and synthetic speech treatments

Four comparison listening effort tests were carried out using four combinations of synthetic and TFW modulated speech. Each test was repeated using a different set of subjects for each repetition giving eight sets of subjective data. The eight sets of data are shown in app. (A.4.5). The analysis of the results for each repetition of the comparison listening effort tests are presented together.

5.5.1 Comparison Listening Effort Test No.1

The first comparison listening effort test used the same natural and synthetic speech treatments as the listening effort test used to determine the TFW parameters. The test was carried out using the procedure described in sec. (4.5), the listening levels used were ---.

Treatments

R1 - Natural speech

S1 - P.D.P.11 Text input synthesizer, variable pitch contour

S2 - P.D.P.11 Text input synthesizer, fixed pitch contour

S3 - Phoneme based system Mac. Talk Phonetic corrections

T1 - TFW Modulation: Sine; Period; 150 ms; Amplitude 0.5 KHz

T2 - TFW Modulation: Sine; Period; 150 ms; Amplitude 3.5 KHz

T3 - TFW Modulation: Sine; Period; 150 ms; Amplitude 3 KHz

T4 - TFW Modulation: Sine; Period; 150 ms; Amplitude 2.5 KHz

T5 - TFW Modulation: Sine; Period; 150 ms; Amplitude 2 KHz

T6 - TFW Modulation: Sine; Period; 150 ms; Amplitude 1.5 KHz

T7 - TFW Modulation: Sine; Period; 150 ms; Amplitude 1 KHz

5.5.1.1 Results of First Comparison Listening Effort Test

The mean listening effort scores over the five listening levels for each treatment were :-

First Repetition

	R1	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
Mean	3.73	1.58	1.36	1.82	3.44	0.22	0.4	0.72	0.75	1.43	2.51

Second Repetition

	R1	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
Mean	3.53	1.25	1.07	1.71	2.98	0.24	0.24	0.42	0.64	1.51	2.4

The mean scores are based on eleven subjects scores for five listening levels, i.e. 55 scores per treatment. The treatments were placed in rank order and plotted as column graphs, figs. (5.19a-b).

Column Graphs ofComparison Listening Effort Test 1 Mean Scores Repetitions1 & 2

Comparison Listening Effort Test 1 Mean Scores Repetition 1

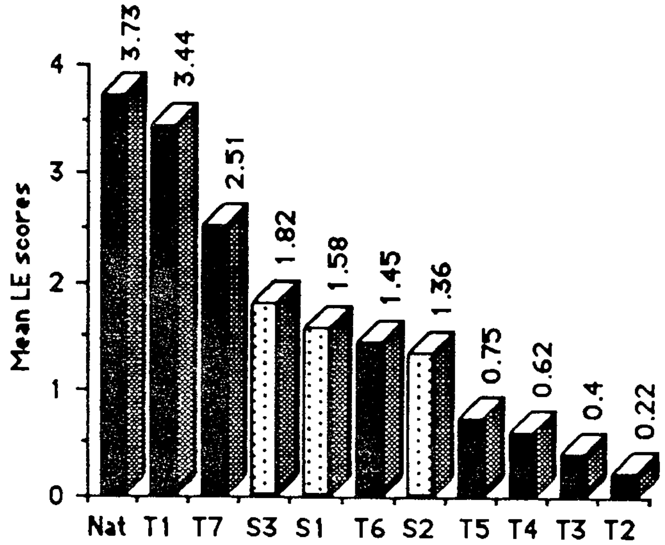


Figure (5.19a)

Comparison Listening Effort Test 1 Mean Scores Repetition 2

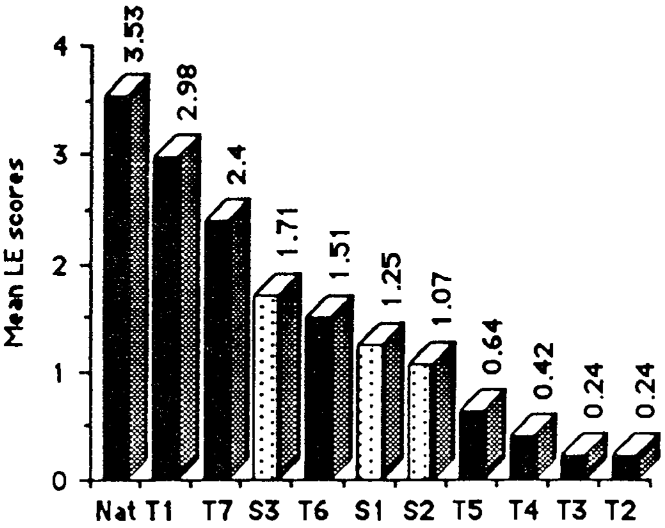
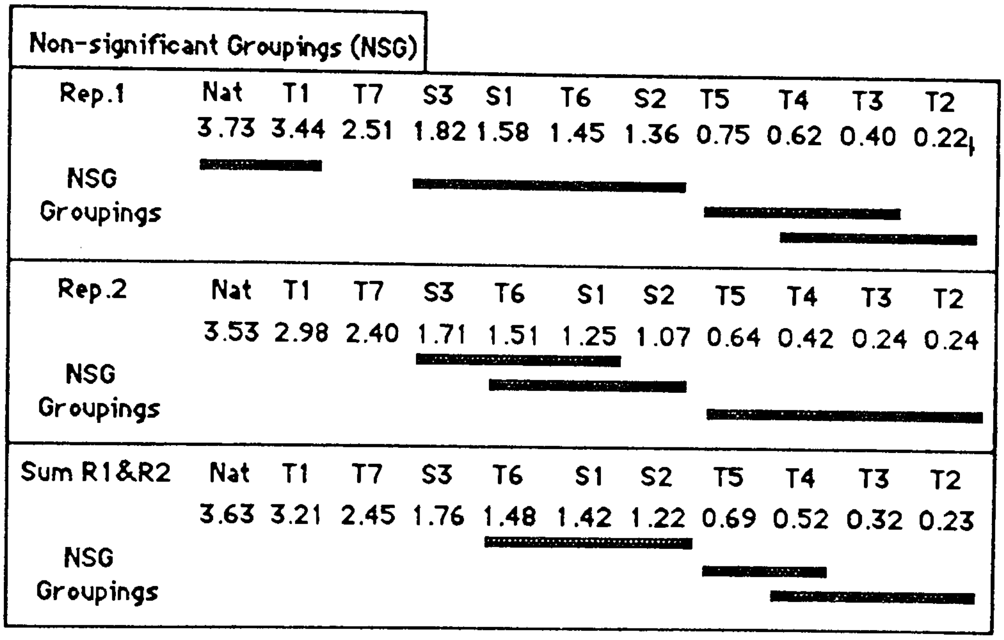


Figure (5.19b)

5.5.1.2 Groupings of Not Significantly Different Treatments

It can be seen from the column graphs figs. (5.19a-b), that except for treatments S1 and T6 the rank order of the treatments remains the same for both repetitions of the comparison listening effort test. To determine whether the difference between S1 and T6 is significant. Plus determine groupings of treatments that are not significantly different. The treatment means were placed in rank order and the F-ratio between all combinations of treatment pairs were calculated. Treatment means that were not significantly different i.e. the F-ratio is below the critical value, are grouped together. 0ure (5.21) shows the grouping of treatments that are not significantly different.

Comparison LE Test 1 NSG of Treatment Mean Scores



Non-significant Groupings (NSG)

Rep.1	Nat	T1	T7	S3	S1	T6	S2	T5	T4	T3	T2
	3.73	3.44	2.51	1.82	1.58	1.45	1.36	0.75	0.62	0.40	0.22
NSG Groupings	▬	▬		▬	▬	▬	▬				
								▬	▬	▬	
									▬	▬	▬
Rep.2	Nat	T1	T7	S3	T6	S1	S2	T5	T4	T3	T2
	3.53	2.98	2.40	1.71	1.51	1.25	1.07	0.64	0.42	0.24	0.24
NSG Groupings				▬	▬	▬					
					▬	▬	▬				
								▬	▬	▬	▬
Sum R1&R2	Nat	T1	T7	S3	T6	S1	S2	T5	T4	T3	T2
	3.63	3.21	2.45	1.76	1.48	1.42	1.22	0.69	0.52	0.32	0.23
NSG Groupings					▬	▬	▬				
								▬	▬		
									▬	▬	▬

Figure (5.20)

Treatments S1 and T6 were not significantly different for both repetitions of the test. Therefore the change in rank order is not significant. The change in rank order may be due to other sources of variance in the data. A complete analysis of variance of the subjective data from the comparison listening effort tests was carried out. The analysis tests whether the main effects and or the interactions are significantly large.

5.5.1.3 Complete Graeco Latin Square x Layers Analysis of Variance

Comparison Listening Effort Test No.1 Rep.1

Factor	d.f.	SS	MS	F-ratio	Sig
Subjects	10	99.43	9.94	9.408	***
Sequences	10	10.23	1.02	0.968	NS
Treatments	10	763.72	76.37	72.260	***
Lists	10	8.70	1.02	0.823	NS
Error(1)	80	84.55	1.06	3.068	***
Levels	4	18.27	4.57	13.261	***
Interactions					
Subjects x Listening Level	40	25.94	0.65	1.883	**
Sequences x Listening Levels	40	16.05	0.40	1.165	NS
Treatments x Listening Levels	40	24.56	0.61	1.782	***
Lists x Listening Levels	40	16.13	0.40	1.170	NS
Error(2)	320	110.24	0.34		
Total	604	1177.83			

Comparison Listening Effort Test No.1 Rep.2

Factor	d.f.	SS	MS	F-ratio	Sig
Subjects	10	60.16	6.02	5.133	***
Sequences	10	9.43	0.94	0.805	NS
Treatments	10	686.82	68.68	58.597	***
Lists	10	20.13	0.94	1.717	NS
Error(1)	80	93.77	1.17	2.622	***
Levels	4	26.49	6.62	14.811	***
Interactions					
Subjects x Listening Level	40	21.08	0.53	1.179	NS
Sequences x Listening Levels	40	21.26	0.53	1.189	NS
Treatments x Listening Levels	40	19.33	0.48	1.081	NS
Lists x Listening Levels	40	16.39	0.41	0.916	NS
Error(2)	320	143.06	0.45		
Total	604	1117.91			

5.5.1.4 Determining TFW Equivalent Listening Effort Scores TFWe

TFWe is the Time Frequency Warping modulation setting that would produce a mean listening effort score equivalent to the synthesized speech mean score sec. (4.5.1). A regression line was calculated based on the TFW treatments mean listening effort scores, i.e. the mean of the five listening levels. The mean scores were transformed using a logistic transformation equation sec. (4.4.3). A y -on- x linear regression lines was calculated and fitted to the data. The graphs for repetition one and two are shown below figs. (5.21a-b) They show the fitted y -on- x regression line and the TFW mean listening effort scores on which it is based. The synthetic speech treatments mean listening effort scores are also shown. The synthetic treatments scores are represented as straight lines on the regression graph.

Y -on- X Regression Lines Comparison Listening Effort Test No.1 Reps.1 & 2

Y -on- X Regression Line Rep.1

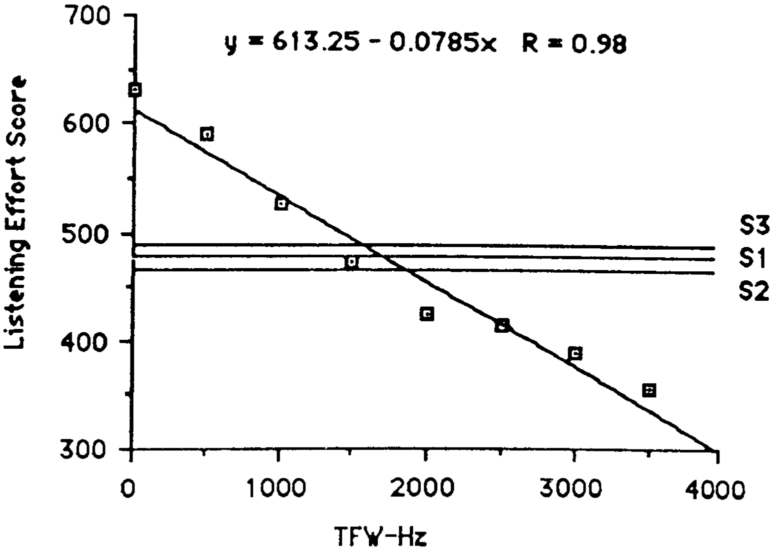


Figure (5.21a)

Y -on- X Regression Line Rep.2

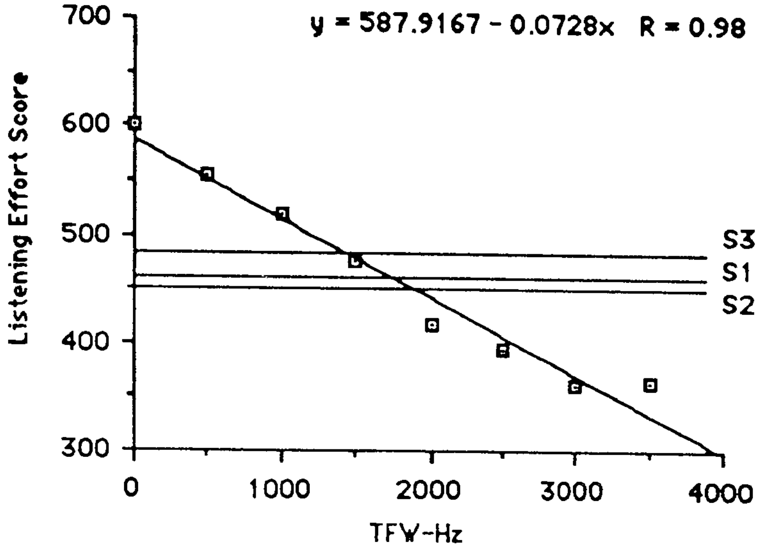


Figure (5.21b)

The intersection of the regression line and the synthetic speech lines determine the TFWe values. The three synthesis systems were;

S1 - P.D.P.11 Text input synthesizer, variable pitch contour

S2 - P.D.P.11 Text input synthesizer, fixed pitch contour

S3 - Phoneme based system Mac. Talk Phonetic corrections

For each synthesis system the mean score over the five listening levels was transformed and used in the calculations.The TFWe of the synthetic treatments mean scores and their confidence limits were calculated sec. (4.5.1) and are shown below.

Repetition 1

	Transformed Listening Effort Scores			TFW Equivalent - TFWe (Hz)		
Treatment	S1	S2	S3	S1	S2	S3
Upper Confidence Limit	493	480	506	1528	1694	1363
Mean Score	479	467	491	1707	1859	1554
Lower Confidence Limit	463	453	475	1910	2058	1757
Repetition 2						
Upper Confidence Limit	474	464	499	1552	1689	1208
Mean Score	461	450	485	1730	1881	1401
Lower Confidence Limit	446	433	471	1936	2115	1593

5.5.1.5 Consistency of TFW Mean Listening Effort Scores

Three TFW settings were selected and the listening effort scores calculated from the y -on- x regression line. The regression line listening effort values for each TFW setting was used because the regression line is based on several sample means. Which gives a more accurate estimate of the "true mean" than the value estimated from the single sample mean for each TFW setting.

TFW Settings Hz	500	1500	2500
LE Score Rep.1	573	495	416
LE Score Rep.2	547	477	405

These results are compared to the equivalent TFW setting used in the other comparison listening effort tests in app.(A.4.5).

5.5.2 Comparison Listening Effort Test No.2

Treatments

R1 - Natural speech

S1 - P.D.P.11 Text input synthesizer, variable pitch contour

S2 - INFOVOX 850905 "Preliminary British"

S3 - KLATT System

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 3 KHz

5.5.2.1 Results of Second Comparison Listening Effort Test

The mean listening effort scores over the five listening levels for each treatment were :-

First Repetition

	R1	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
Mean	3.81	1.15	1.11	1.11	3.61	3.2	3.2	2.69	1.85	0.8	0.1

Second Repetition

	R1	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
Mean	3.35	1.24	1.27	1.05	3.2	3.09	3.09	2.52	1.32	0.65	0.12

Mean scores are based on eleven subjects scores for five listening levels, i.e. 55 scores per treatment. The treatments were placed in rank order and plotted as column graphs, figs. (5.22a-b).

Column Graphs of Comparison Listening Effort Test 2 Mean Scores

Comparison Listening Effort Test 2 Mean Scores Repetition 1

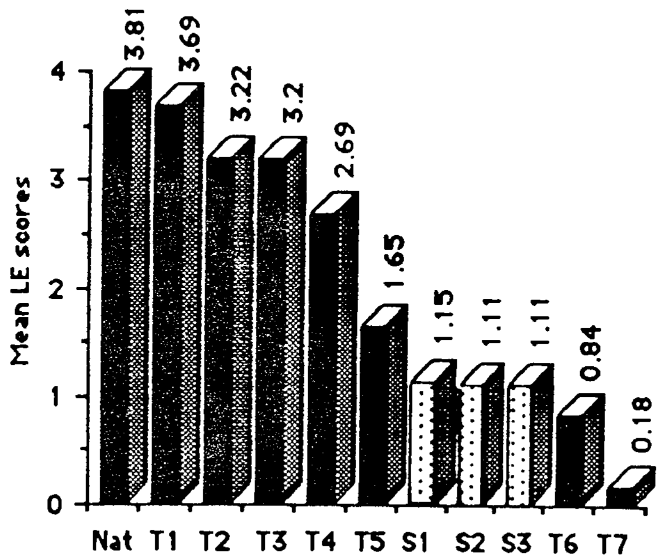


Figure (5.22a)

Comparison Listening Effort Test Mean Scores Repetition 2

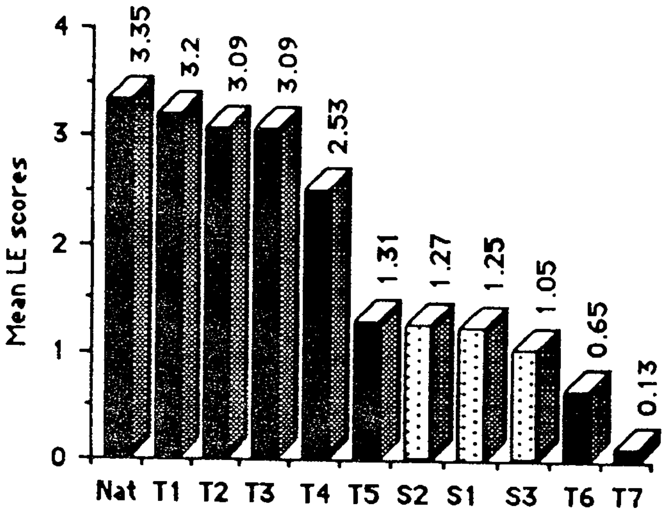
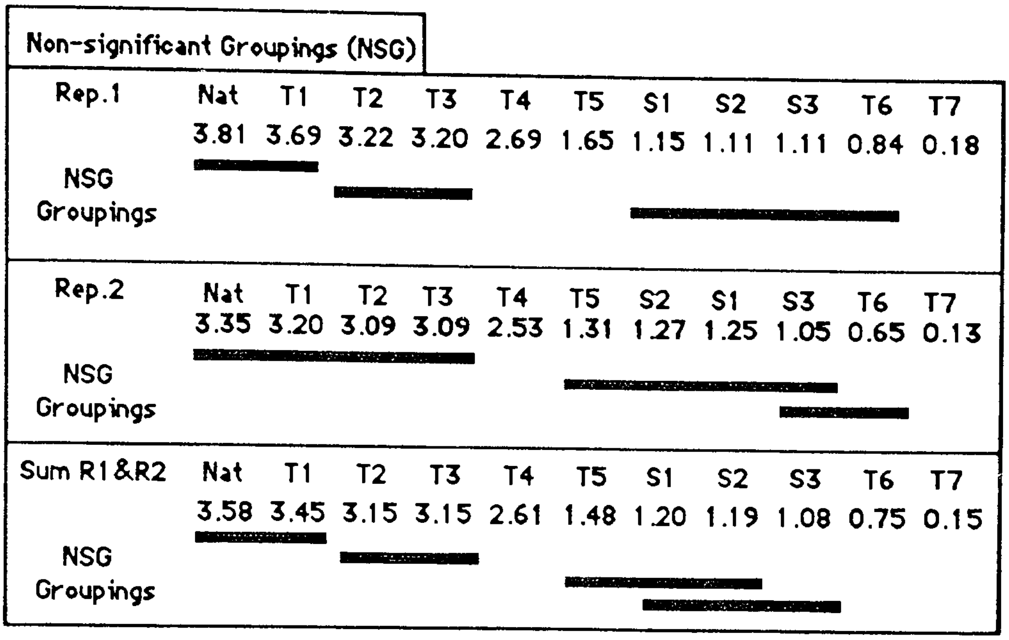


Figure (5.22b)

5.5.2.2 Groupings of Not Significantly Different Treatments

It can be seen from the column graphs figs. (5.22a-b), that except for treatments S1 and S2 the rank order of the treatments remains the same for both repetitions of the comparison listening effort tests. To determine the treatments that are not significantly different. The treatment means were placed in rank order and the F-ratio between all combinations of treatment pairs were calculated. Treatment means that were not significantly different i.e. the F-ratio is below the critical value, are grouped together. Figure (5.23) shows the grouping of treatments that are not significantly different.

Comparison LE Test 2 NSG of Treatment Mean Scores



Non-significant Groupings (NSG)

Rep.1	Nat	T1	T2	T3	T4	T5	S1	S2	S3	T6	T7
	3.81	3.69	3.22	3.20	2.69	1.65	1.15	1.11	1.11	0.84	0.18

NSG Groupings

Rep.2	Nat	T1	T2	T3	T4	T5	S2	S1	S3	T6	T7
	3.35	3.20	3.09	3.09	2.53	1.31	1.27	1.25	1.05	0.65	0.13

NSG Groupings

Sum R1&R2	Nat	T1	T2	T3	T4	T5	S1	S2	S3	T6	T7
	3.58	3.45	3.15	3.15	2.61	1.48	1.20	1.19	1.08	0.75	0.15

NSG Groupings

Figure (5.23)

Synthetic treatments S1 and S2 mean scores are not significantly different from each other in both repetitions. Therefore the change in rank order of S1 and S2 is not significant. The change in rank order may be due to other sources of variance in the data. A complete analysis of variance of the subjective data from the comparison listening effort tests was carried out. The analysis tests whether the main effects and or the interactions are significantly large.

5.5.2.3 Complete Graeco Latin Square x Layers Analysis of Variance

Comparison Listening Effort Test No.2 Rep. 1

Factor	d.f.	SS	MS	F-ratio	Sig
Subjects	10	79.71	7.97	9.893	***
Sequences	10	8.69	0.87	1.079	NS
Treatments	10	914.33	91.43	113.47	***
Lists	10	10.66	1.07	1.323	NS
Error(1)	80	64.46	0.81	2.44	***
Levels	4	26.27	6.57	19.89	***

Interactions					
Subjects x Listening Level	40	28.06	0.70	2.142	**
Sequences x Listening Levels	40	11.62	0.29	0.880	NS
Treatments x Listening Levels	40	14.35	0.36	1.086	NS
Lists x Listening Levels	40	14.02	0.35	1.061	NS
Error(2)	320	105.61	0.33		
Total	604	1277.06			

Comparison Listening Effort Test No.2 Rep. 2

Factor	d.f.	SS	MS	F-ratio	Sig
Subjects	10	42.92	4.29	2.762	**
Sequences	10	10.08	1.01	0.642	NS
Treatments	10	746.74	74.67	48.047	***
Lists	10	21.97	2.20	1.414	NS
Error(1)	80	124.33	1.55	4.838	***
Levels	4	38.47	9.62	24.94	***
Interactions					
Subjects x Listening Level	40	17.86	0.45	1.390	NS
Sequences x Listening Levels	40	15.97	0.40	1.243	NS
Treatments x Listening Levels	40	21.86	0.55	1.701	**
Lists x Listening Levels	40	16.26	0.41	1.265	NS
Error(2)	320	102.79	0.32		
Total	604	1159.25			

5.5.2.4 Determining TFW Equivalent Listening Effort Scores TFWe

A regression line was calculated based on the TFW treatments mean listening effort scores, i.e. the mean of the five listening levels. The mean scores were transformed using a logistic transformation equation sec. (4.5.1). The y -on- x linear regression line was calculated and fitted to the data. The graphs for repetition one and two are shown below fig. (5.24a-b) They show the fitted y -on- x regression line and the TFW listening effort mean scores on which it is based. The synthetic speech treatments mean scores are also shown. The synthetic treatments mean listening effort scores are represented as straight lines on the regression graph.

Y -on- X Regression Lines Comparison Listening Effort Test No.2 Reps.1 & 2

Y -on- X Regression Rep 1

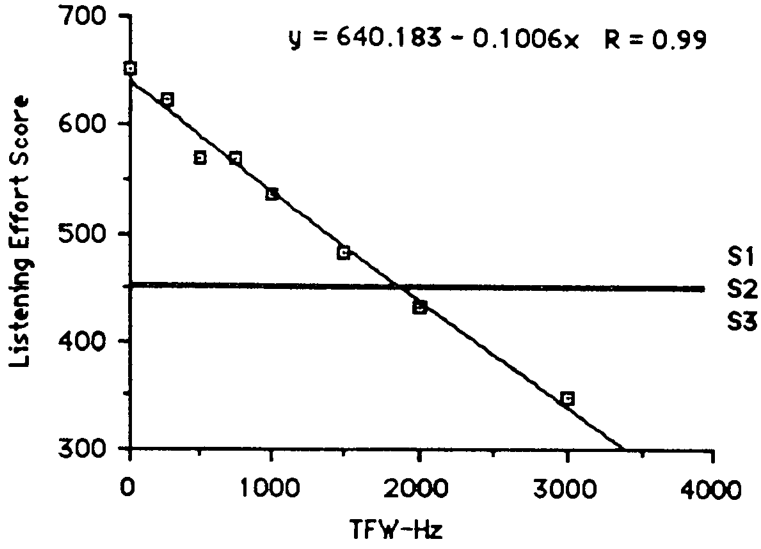


Figure (5.24a)

Y -on- X Regression Line Rep. 2

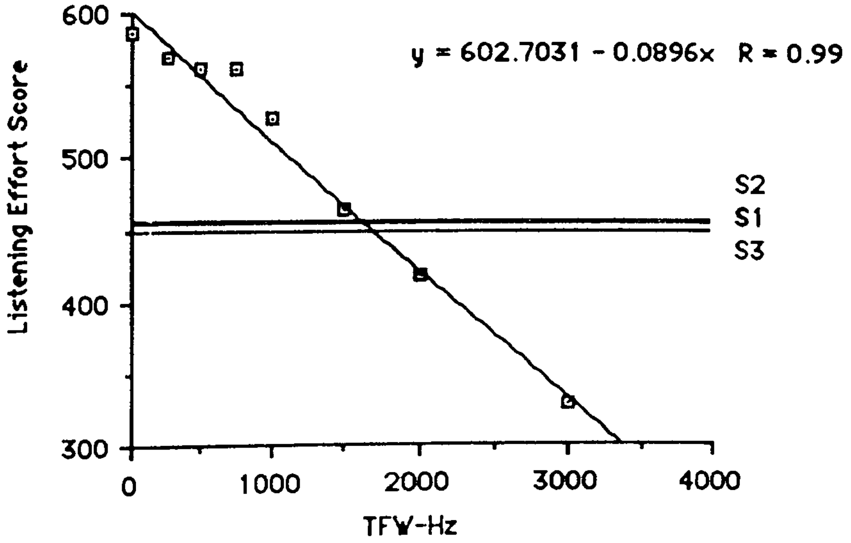


Figure (5.24b)

The intersection of the regression line and the synthetic speech lines determine the TFWe values. The three synthesis systems were;

S1 - P.D.P.11 Text input synthesizer, variable pitch contour

S2 - INFOVOX 850905 "Preliminary British"

S3 - KLATT System

For each synthesis system the mean score over the five levels was transformed and used in the calculations.The TFWe of the synthetic treatments mean scores and their confidence limits were calculated and are shown below.

Repetition 1

	Transformed Listening Effort			TFW Equivalent - TFWe (Hz)		
Treatment	S1	S2	S3	S1	S2	S3
Upper Confidence Limit	469	466	466	1698	1729	1729
Mean Score	454	452	452	1848	1868	1868
Lower Confidence Limit	437	436	436	2027	2017	2017

Repetition 2

Treatment	S1	S2	S3	S1	S2	S3
Upper Confidence Limit	473	475	460	1439	1417	1584
Mean Score	461	462	449	1573	1562	1707
Lower Confidence Limit	447	448	436	1729	1718	1852

5.5.2.5 Consistency of TFW Mean Listening Effort Scores

Three TFW settings were selected and the listening effort scores calculated from the y -on- x regression line.

TFW Settings Hz	500	1500	2500
LE Score Rep.1	589	489	388
LE Score Rep.2	557	467	378

5.5.3 Comparison Listening Effort Test No.3

Treatments

R1 - Natural speech

S1 - INFOVOX American version

S2 - INFOVOX 850905 "Preliminary British"

S3 - INFOVOX British (From Sweden)

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 2.5 KHz

T7 - TFW Modulation: Sine; Period; 150 ms; Amplitude 2 KHz

5.5.3.1 Results of Comparison Listening Effort Test No.3

The mean listening effort scores over the five listening levels for each treatment were :-

First Repetition

	R1	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
Mean	3.64	1.58	1.27	1.53	3.6	3.02	3.3	2.64	1.18	0.56	1.09

Second Repetition

	R1	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
Mean	3.78	2.0	1.42	1.93	3.58	3.3	3.3	2.8	1.49	0.65	1.05

Mean scores are based on eleven subjects scores for five listening levels, i.e. 55 scores per treatment. The treatments were placed in rank order and plotted as a column graphs, figs. (5.25a-b).

Column Graphs of Comparison Listening Effort Test 3 Mean Scores

Comparison Listening Effort Test 3 Mean Scores Repetition 1

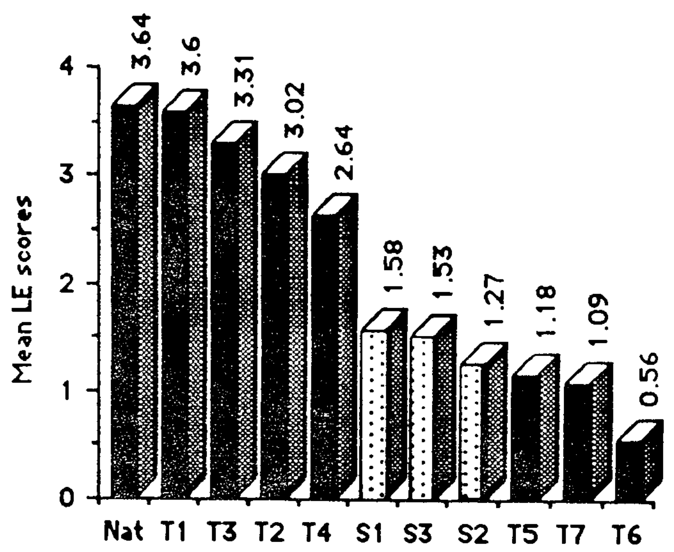


Figure (5.25a)

Comparison Listening Effort Test 3 Mean Scores Repetition 2

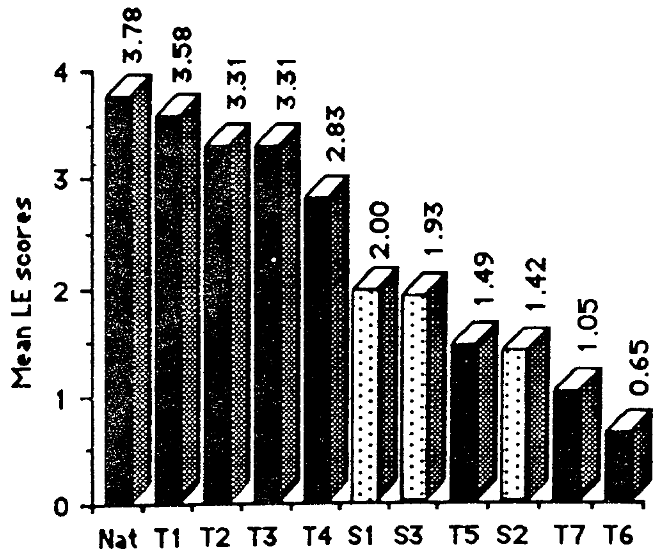
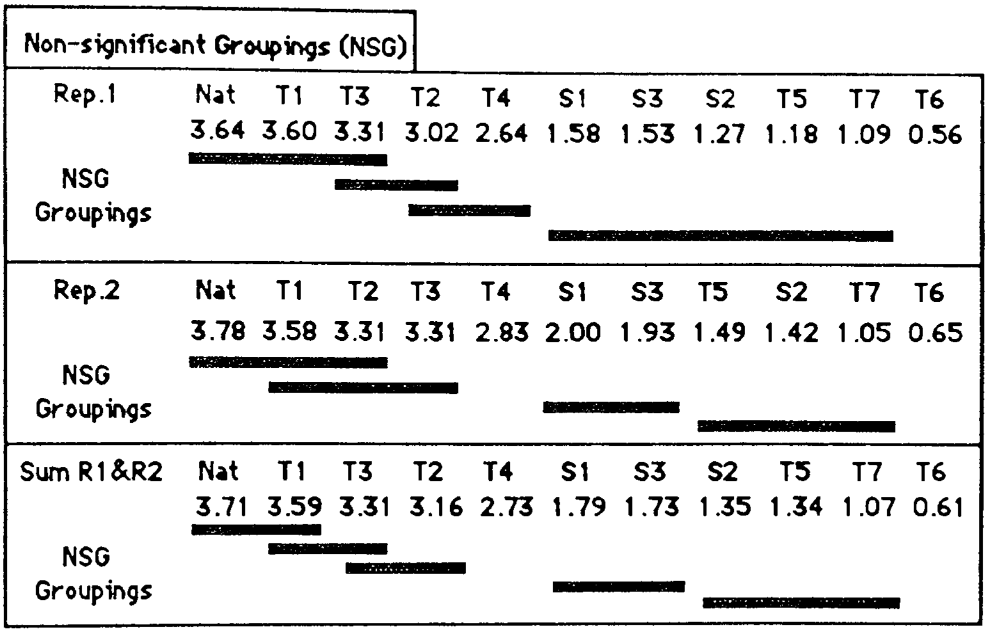


Figure (5.25b)

5.5.3.2 Groupings of Not Significantly Different Treatments

It can be seen from the column graphs figs. (5.25a-b), that except for treatments T5 and S2 the rank order of the treatments remains the same for both repetitions of the comparison listening effort tests. To determine the treatments that are not significantly different. The treatment means were placed in rank order and the F-ratio between all combinations of treatment pairs were calculated. Treatment means that were not significantly different i.e. the F-ratio is below the critical value, are grouped together. Figure (5.26) shows the grouping of treatments that are not significantly different.

Comparison LE Test 3 NSG of Treatment Mean Scores



Non-significant Groupings (NSG)

Rep.1	Nat	T1	T3	T2	T4	S1	S3	S2	T5	T7	T6
	3.64	3.60	3.31	3.02	2.64	1.58	1.53	1.27	1.18	1.09	0.56
NSG Groupings											

Rep.2	Nat	T1	T2	T3	T4	S1	S3	T5	S2	T7	T6
	3.78	3.58	3.31	3.31	2.83	2.00	1.93	1.49	1.42	1.05	0.65
NSG Groupings											

Sum R1&R2	Nat	T1	T3	T2	T4	S1	S3	S2	T5	T7	T6
	3.71	3.59	3.31	3.16	2.73	1.79	1.73	1.35	1.34	1.07	0.61
NSG Groupings											

Figure (5.26)

Synthetic treatments T5 and S2 mean scores are not significantly different from each other in both repetitions. Therefore the change in rank order of T5 and S2 is not significant. The change in rank order may be due to other sources of variance in the data. A complete analysis of variance of the subjective data from the comparison listening effort tests was carried out. The analysis tests whether the main effects and or the interactions are significantly large.

5.5.3.3 Complete Graeco Latin Square x Layers Analysis of Variance

Comparison Listening Effort Test No.3 Rep. 1

Factor	d.f.	SS	MS	F-ratio	Sig
Subjects	10	64.09	6.41	4.724	***
Sequences	10	19.54	1.95	1.441	NS
Treatments	10	698.31	69.83	51.47	***
Lists	10	17.07	1.71	1.258	NS
Error(1)	80	108.53	1.36	2.815	***
Levels	4	22.49	5.62	11.67	***
Interactions					
Subjects x Listening Level	40	30.82	0.77	1.599	*
Sequences x Listening Levels	40	11.73	0.29	0.609	NS
Treatments x Listening Levels	40	20.24	0.51	1.050	NS
Lists x Listening Levels	40	18.93	0.47	0.982	NS
Error(2)	320	154.20	0.48		
Total	604	1165.94			

Comparison Listening Effort Test No.3 Rep. 2

Factor	d.f.	SS	MS	F-ratio	Sig
Subjects	10	103.35	10.33	9.904	***
Sequences	10	7.35	0.73	0.704	NS
Treatments	10	663.49	66.35	63.58	***
Lists	10	11.57	1.16	1.108	NS
Error(1)	80	83.48	1.04	2.346	***
Levels	4	9.38	2.34	5.271	***
Interactions					
Subjects x Listening Level	40	25.31	0.63	1.423	NS
Sequences x Listening Levels	40	25.31	0.63	1.423	NS
Treatments x Listening Levels	40	16.08	0.40	0.903	NS
Lists x Listening Levels	40	24.37	0.61	1.369	NS
Error(2)	320	142.35	0.44		
Total	604	1112.04			

5.5.3.4 Determining TFW Equivalent Listening Effort Scores TFWe

A regression line was calculated based on the TFW treatments mean listening effort scores, i.e. the mean of the five listening levels. The mean scores were transformed using a logistic transformation equation sec. (4.5.1). The y -on- x linear regression line was calculated and fitted to the data. The graphs for repetition one and two are shown below figs. (5.27a-b) They show the fitted y -on- x regression line and the TFW listening effort mean scores on which it is based. The synthetic speech treatments mean scores are also shown. The synthetic treatments mean listening effort scores are represented as straight lines on the regression graph.

Y -on- X Regression Lines Comparison Listening Effort Test No.3 Reps.1 & 2

Y -on- X Regression Line Repetition 1

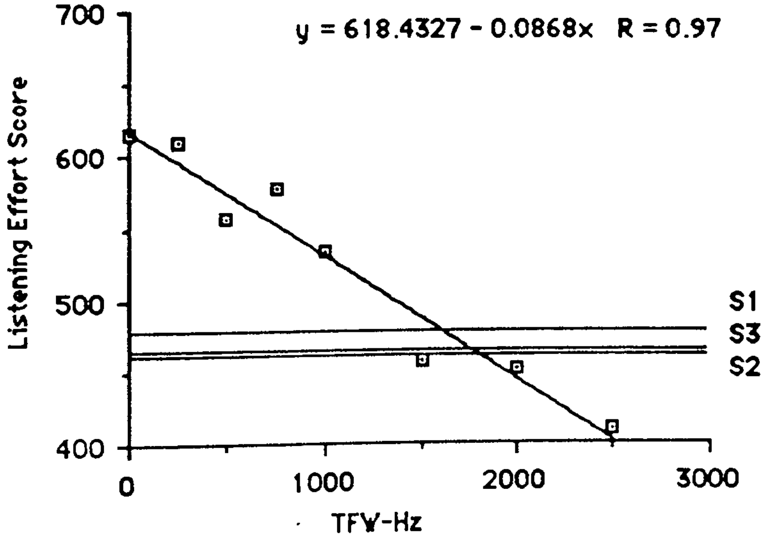


Figure (5.27a)

Y -on- X Regression Line Repetition 2

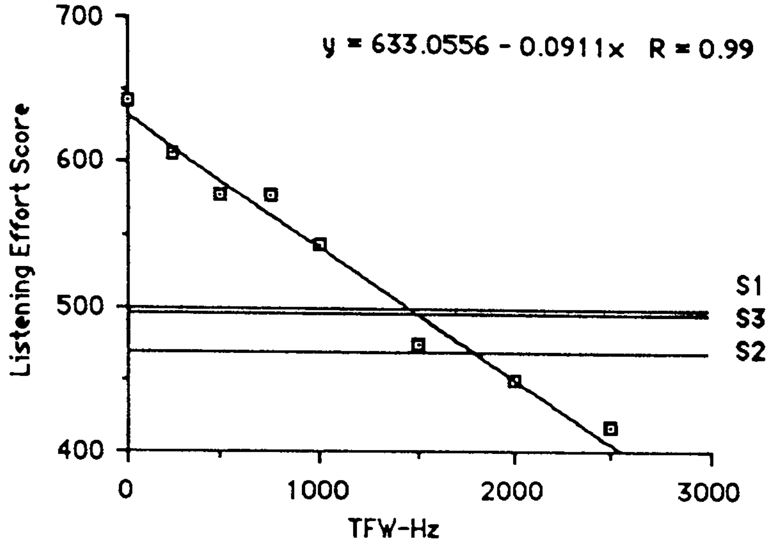


Figure (5.27b)

The intersection of the regression line and the synthetic speech lines determine the TFWe values.

The three synthesis systems were;

S1 - INFOVOX American version

S2 - INFOVOX 850905 "Preliminary British"

S3 - INFOVOX British (From Sweden)

For each synthesis system the mean score over the five levels was transformed and used in the calculations.The TFWe of the synthetic treatments mean scores and their confidence limits were calculated and are shown below.

Repetition 1

	Transformed Listening Effort			TFW Equivalent - TFWe		
Treatment	S1	S2	S3	S1	S2	S3
Upper Confidence Limit	492	477	491	1451	1624	1463
Mean Score	479	462	476	1601	1797	1635
Lower Confidence Limit	465	445	460	1762	1992	1820

Repetition 2

Upper Confidence Limit	514	485	510	1306	1624	1350
Mean Score	500	470	496	1460	1789	1503
Lower Confidence Limit	486	454	482	1613	1964	1657

5.5.3.5 Consistency of TFW Mean Listening Effort Scores

Three TFW settings were selected and the listening effort scores calculated from the y -on- x regression line.

TFW Settings Hz	500	1500	2500
LE Score Rep.1	575	487	401
LE Score Rep.2	585	496	405

5.5.4 Comparison Listening Effort Test No.4

Treatments

R1 - Natural speech

S1 - INFOVOX American version

S2 - INFOVOX 850905 "Preliminary British"

S3 - DEC Talk American

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 2.5 KHz

T7 - TFW Modulation: Sine; Period; 150 ms; Amplitude 2 KHz

5.5.4.1 Results of Comparison Listening Effort Test No.4

The mean listening effort scores over the five listening levels for each treatment were :-

	R1	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
First Repetition											
Mean	3.78	1.67	1.29	2.85	3.56	3.25	3.13	2.36	1.53	0.64	1.02
Second Repetition											
Mean	3.64	1.62	1.45	2.58	3.45	3.2	3.04	2.67	1.45	0.65	1.25

Mean scores are based on eleven subjects scores for five listening levels, i.e. 55 scores per treatment. The treatments were placed in rank order and plotted as a column graphs, figs. (5.28a-b).

Column Graphs of Comparison Listening Effort Test 4 Mean Scores

Comparison Listening Effort Test 4 Mean Scores Repetition 1

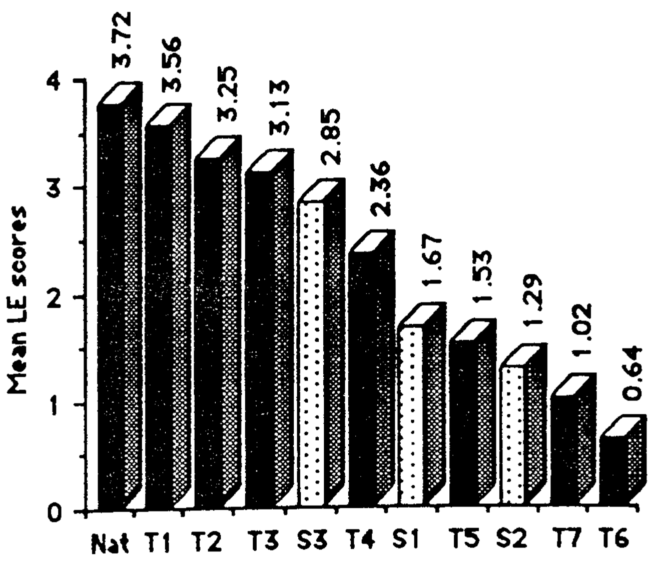


Figure (5.28a)

Comparison Listening Effort Test 4 Mean Scores Repetition 2

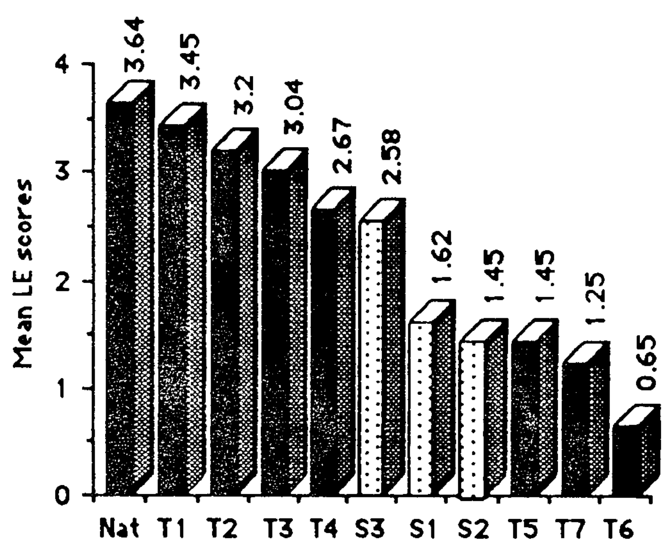


Figure (5.28b)

5.5.4.2 Groupings of Not Significantly Different Treatments

It can be seen from the column graphs figs. (5.28a-b), treatments T5 & S2 and S3 & T4 changed places in the rank order. To clarify the changes in rank order and determine the groupings of treatments that are not significantly different. The treatment means were placed in rank order and the F-ratio between all combinations of treatment pairs were calculated. Treatment means that were not significantly different i.e. the F-ratio is below the critical value, are grouped together. Figure (5.29) shows the grouping of treatments that are not significantly different.

Comparison LE Test 4 NSG of Treatment Mean Scores

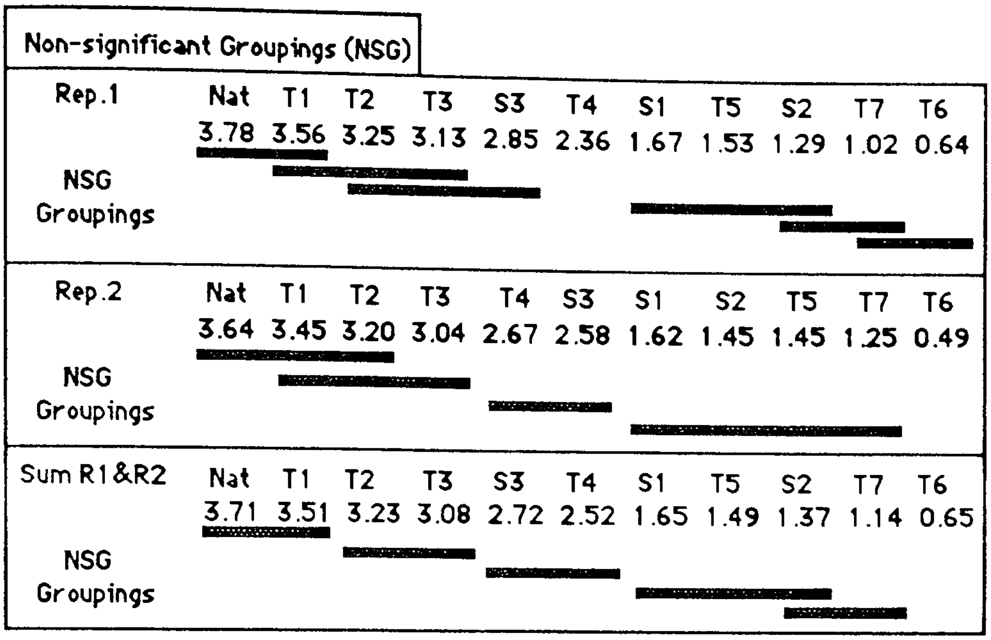


Figure (5.29)

Synthetic treatments T5 and S2 mean scores are not significantly different from each other in both repetitions. Therefore the change in rank order of T5 and S2 is not significant. Whereas S3 and T4 are significantly different in the first repetition but not in the second. The two treatments change rank order in the second repetition. The change in rank order may be due to other sources of variance in the data. A complete analysis of variance of the subjective data from the comparison listening effort tests was carried out. The analysis tests whether the main effects and or the interactions are significantly large.

5.5.4.3 Complete Graeco Latin Square x Layers Analysis of Variance

Comparison Listening Effort Test No.4 Rep. 1

Factor	d.f.	SS	MS	F-ratio	Sig
Subjects	10	57.21	5.72	5.393	***
Sequences	10	8.12	0.81	0.766	NS
Treatments	10	666.34	66.63	62.81	***
Lists	10	14.89	1.49	1.403	NS
Error(1)	80	84.87	1.06	2.419	***
Levels	4	28.08	7.02	16.011	***

Interactions					
Subjects x Listening Level	40	26.57	0.66	1.515	*
Sequences x Listening Levels	40	13.84	0.35	0.789	NS
Treatments x Listening Levels	40	21.99	0.55	1.254	NS
Lists x Listening Levels	40	9.99	0.25	0.570	NS
Error(2)	320	140.32	0.44		
Total	140.32		0.44		

Comparison Listening Effort Test No.4 Rep. 2

Factor	d.f.	SS	MS	F-ratio	Sig
Subjects	10	137.73	13.77	14.93	***
Sequences	10	9.91	0.99	1.074	NS
Treatments	10	570.74	57.07	61.86	***
Lists	10	46.67	4.67	5.059	***
Error(1)	80	73.80	0.92	2.650	***
Levels	4	20.92	5.23	15.02	***
Interactions					
Subjects x Listening Level	40	31.08	0.78	2.232	*
Sequences x Listening Levels	40	12.36	0.31	0.887	NS
Treatments x Listening Levels	40	10.98	0.27	0.788	NS
Lists x Listening Levels	40	8.87	0.22	0.637	NS
Error(2)	320	111.40	0.35		
Total	604	1034.45			

5.5.4.4 Determining TFW Equivalent Listening Effort Scores TFWe

A regression line was calculated based on the TFW treatments mean listening effort scores, i.e. the mean of the five listening levels. The mean scores were transformed using a logistic transformation equation sec. (4.5.1). The y -on- x linear regression line was calculated and fitted to the data. The graphs for repetition one and two are shown below fig. (5.30a-b) They show the fitted y -on- x regression line and the TFW listening effort mean scores on which it is based. The synthetic speech treatments mean scores are also shown. The synthetic treatments mean listening effort scores are represented as straight lines on the regression graph.

Y -on- X Regression Lines Comparison Listening Effort Test No.4 Reps.1 & 2

Y -on- X Regression Line Repetition 1

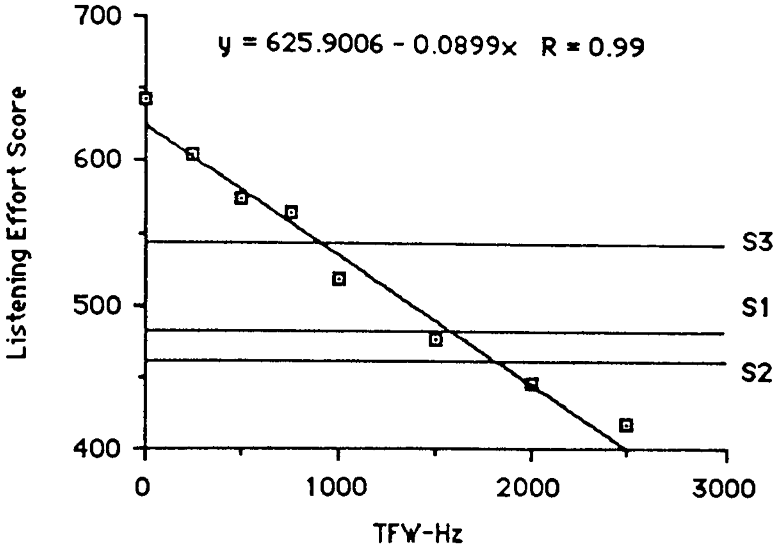


Figure (5.30a)

Y -on- X Regression Line Repetition 2

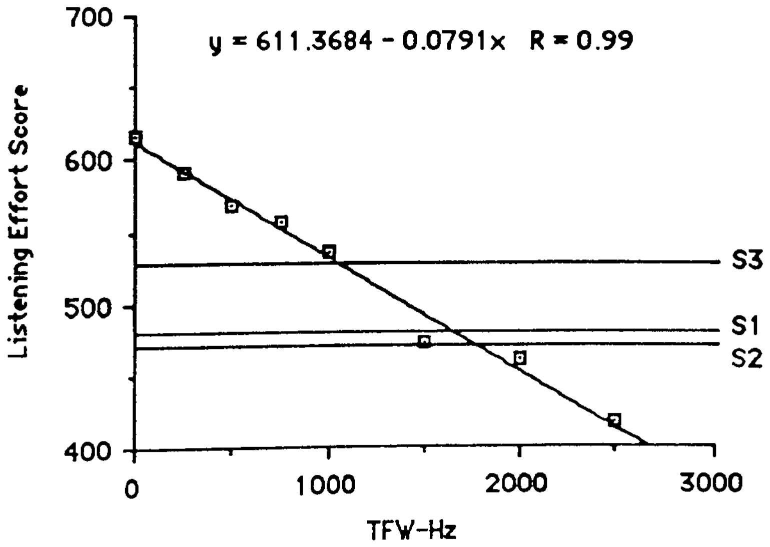


Figure (5.30b)

Determining Synthetic Speech Treatments TFW Equivalents TFWe

The intersection of the regression line and the synthetic speech lines determine the TFWe values. The three synthesis systems were;

S1 - INFOVOX American version

S2 - INFOVOX 850905 "Preliminary British"

S3 - DEC Talk American

For each synthesis system the mean score over the five levels was transformed and used in the calculations.The TFWe of the synthetic treatments mean scores and their confidence limits were calculated and are shown below.

	Transformed Listening Effort			TFW Equivalent - TFWe		
Treatment	S1	S2	S3	S1	S2	S3
Upper Confidence Limit	495	476	558	1446	1657	745
Mean Score	483	463	546	1579	1802	878
Lower Confidence Limit	471	449	534	1713	1957	1019
Repetition 2						
Upper Confidence Limit	494	486	546	1479	1580	821
Mean Score	481	472	530	1643	1757	1024
Lower Confidence Limit	467	456	516	1820	1959	1201

5.5.4.5 Consistency of TFW Mean Listening Effort Scores

Three TFW settings were selected and the listening effort scores calculated from the y -on- x regression line.

TFW Settings Hz	500	1500	2500
LE Score Rep.1	580	490	400
LE Score Rep.2	571	492	413

5.5.5 Comparison of TFW Mean Listening Effort Scores

For three TFW modulation settings the predicted listening effort scores were calculated from the regression lines of each listening effort test. The purpose of this was to determine the consistency of the predicted listening effort scores over the eight listening effort test. The three TFW modulation settings were 500, 1500, and 2500. The table ( ) below shows the predicted listening effort scores over the eight listening effort tests, (4 x 2 repetitions).

TFW Settings Hz	500	1500	2500
Comparison -LE1 Score Rep.1	573	495	416
Comparison- LE1 Score Rep.2	547	477	405
Comparison- LE2 Score Rep.1	589	489	388
Comparison- LE2 Score Rep.2	557	467	378
Comparison- LE3 Score Rep.1	575	487	401
Comparison- LE3 Score Rep.2	585	496	405
Comparison- LE4 Score Rep.1	580	490	400
Comparison- LE4 Score Rep.2	571	492	413
Mean Predicted Listening Effort Score (P - LE)	572	486	400

The range of predicted scores from which the mean values were calculated, were compared with the regression line confidence limits. The upper (Ucl) and lower (Lcl) confidence limits of the regression lines for Comparison listening effort tests 1 rep. 1, 3 rep. 1, and 4 rep. 2 were calculated. The results are compared in figure (5.31).

Range of Predicted Listening Effort Scores

TFW setting	Mean P- LE Scores	Range P - LE Scores	Regresion Line confidence Limits
500	572	585 547	597 594 580 Ucl 546 553 560 Lcl
1500	486	496 467	513 506 501 Ucl 475 468 482 Lcl
2500	400	416 378	437 435 429 Ucl 394 366 396 Ucl

Figure (5.31)

It can be seen from fig.(5.31) that for TFW settings 1500 and 2500 the range of predicted listening effort scores (P - LE) over the eight tests fall within the regression line confidence limits of the individual tests shown in the table. For TFW setting 500 the P - LE range of scores is greater than the confidence limits of the individual test shown in the table. The largest deviation is 13 points on the transformed listening effort scale. This is approximately 0.2 on the untransformed scale. If this figure 0.2 is compared to the values for the difference between treatment mean scores in the non-significant grouping fig. (5.29) Range of Predicted Listening Effort Scores). It is seen that a difference of 0.2 between treatments is not significant. That is to say that the P - LE scores which fall outside the individual tests confidence limits are not significantly different from P - LE scores that have the same scores as the confidence limits.

5.5.6 Consistency of Synthetic Speech Listening Effort Scores

One synthetic speech treatment "INFOVOX 850905 "Preliminary British" was used in three of the comparison listening effort tests. The listening effort test scores and TFW equivalent scores were compared;

	Listening Effort Score	TFW equivalent TFWe
	Ucl Lcl	Lcl Ucl
CLE2	466 - 452 - 436	1729 - 1868 - 2017
	475 - 462 - 448	1417 - 1562 - 1718
CLE3	477 - 462 - 445	1624 - 1797 - 1963
	485 - 470 - 454	1624 - 1789 - 1964
CLE4	476 - 463 - 449	1657 - 1802 - 1957
	486 - 472 - 456	1580 - 1757 - 1959

Ucl - Upper confidence limit Lcl - lower confidence limit

The listening effort scores for synthesis system "INFOVOX 850905 "Preliminary British" over the six tests (3 x 2 reps.) are similar, with a range from 452 - to - 472 which corresponds to a difference of 0.23 on the untransformed listening effort scale. The range of listening effort scores over the six tests is comparable to the confidence limits of the listening effort scores of the individual tests. Which suggests that the differences in the listening effort scores for different tests is comparable to the differences found within a test.

5.5.7 TFW Equivalents for the Six Synthetic Speech Systems

The listening effort scores of the six text - to - speech synthetic speech systems used in the comparison listening effort tests are shown below with their TFW equivalent (TFWe) settings.

TFW Equivalent (TFWe) settings.

	LE score	TFWe (Hz)
P.D.P.11 Text input synthesizer, variable pitch contour	1.19	1.85
INFOVOX 850905 "Preliminary British"	1.21	1.76
INFOVOX American version	1.74	1.53
INFOVOX British (From Sweden)	1.73	1.56
KLATT System	1.05	1.78
DEC Talk American	2.71	0.95

The speech over a telephone connection must have a listening effort score over 2.5 to be considered for use in the public network. A listening effort score of 2.5 has a TFW equivalent of approximately 1 KHz. Of the systems tested the DEC Talk system is the only one which would be considered for use over public network. But it should be noted that it was tested using a Intermediate Reference System, sec. (2.2.3), which only simulates the bandwidth of a telephone connection and none of the degradations found in the public network. From the listening effort scores of the other synthetic speech systems it can be seen that the quality of speech was quite poor.

5.5.8 Comparison Subject Variance

5.5.8.1Comparison of Treatments Subject Variance

Individual subjects listening effort scores for different types of speech are dependant on the individuals abilities and experiances. For example an airline pilot has no problem understanding the poor quality speech he hears in his headphones. Whereas most people would need alot of effort to understand the speech. This is an extreme example, but it highlights the effects poor quality speech has on individual subject's scores. The effects of differing speech quality and the differences between subjects can be seen as differences in the variance of the mean score for different types of speech. If subjects are rating high quality natural speech you would expect the variance of the mean score to be small indicating a high degree of consistency of subjects scores. It could also be expected that the variance of the mean score for very low quality speech to be low. For speech quality that is somewhere between the two extremes the variance will vary dependant on the quality of the speech and the individual subjects. The variance of the treatment means (11 subjects over 5 listening levels) were calculated. Bartlet's test app. (A.2.5) was used to determine the homogenity of the variances. If the variance for a particular treatment mean is significantly high compared to the other mean variances it indicates that the group of subjects were having trouble consistently rating the speech.

Comparison Listening Effort Test 1

First Repetition

	R1	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
Mean	3.73	1.58	1.36	1.82	3.44	0.22	0.4	0.72	0.75	1.43	2.51
Variance	0.35	1.14	0.83	1.26	0.55	0.28	0.47	0.76	0.49	0.85	0.70

Second Repetition

	R1	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
Variance	0.36	0.82	0.81	1.06	1.06	0.63	0.18	0.43	0.46	0.96	1.21

Homogenity of Variances

	Rep1	Rep2	Critial Values
All Treatments	10.09	13.56	18.31
TFW Treatments	10.73	3.64	12.59
Syn Treatments	0.44	0.34	5.99

Comparison Listening Effort Test 2

First Repetition

	R1	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
Variance	0.19	0.94	0.73	0.77	0.44	0.88	0.42	0.74	0.75	0.69	0.19

Second Repetition

	R1	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
Variance	0.75	0.67	0.72	0.46	1.31	0.97	0.64	0.88	0.70	0.42	0.11

Homogenity of Variances

	Rep1	Rep2	Critial Values
All Treatments	13.07	14.66	18.31
TFW Treatments	6.85	13.84	12.59
Syn Treatments	0.18	0.54	5.99

Comparison Listening Effort Test 3

First Repetition

	R1	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
Variance	0.31	0.91	0.98	1.11	0.43	0.65	0.74	1.01	0.97	0.69	0.86

Second Repetition

	R1	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
Variance	0.21	1.00	1.06	1.03	0.36	0.62	0.59	1.23	0.77	0.56	0.87

Homogenity of Variances

	Rep1	Rep2	Critial Values
All Treatments	6.457	11.30	18.31
TFW Treatments	2.286	4.26	12.59
Syn Treatments	0.08	0.08	5.99

Comparison Listening Effort Test 4

First Repetition

	R1	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
Variance	0.21	0.74	0.77	0.53	0.36	0.93	0.71	0.68	0.98	0.68	0.94

Second Repetition

	R1	S1	S2	S3	T1	T2	T3	T4	T5	T6	T7
Variance	0.27	0.91	1.03	1.03	0.59	0.87	0.78	0.85	0.62	0.49	1.16

Homogenity of Variances

	Rep1	Rep2	Critial Values
All Treatments	8.36	7.23	18.31
TFW Treatments	2.98	2.42	12.59
Syn Treatments	0.37	0.05	5.99

5.5.9 Modulated Noise Reference Unit (MNRU) Listening Effort Test

The modulated noise reference unit is a reference device used by British Telecom to assess the effects of digital devices used in the telephone network. The MNRU simulates the effects of quantization noise introduced by some digital devices, ref.( ). This experiment compares the listening effort scores for five settings of the modulated noise reference unit with five settings of the proposed reference device TFW modulation.

Treatments

R1 - Natural speech

M1 - MNRU signal to noise ratio 20 dB

M2 - MNRU signal to noise ratio 15 dB

M3 - MNRU signal to noise ratio 10 dB

M4 - MNRU signal to noise ratio 5 dB

M5 - MNRU signal to noise ratio 0 dB

T1 - TFW Modulation: Sine; Period; 150 ms; Amplitude 0.5 KHz

T2 - TFW Modulation: Sine; Period; 150 ms; Amplitude 1 KHz

T3 - TFW Modulation: Sine; Period; 150 ms; Amplitude 1.5 KHz

T4 - TFW Modulation: Sine; Period; 150 ms; Amplitude 2 KHz

T5 - TFW Modulation: Sine; Period; 150 ms; Amplitude 2.5 KHz

5.5.9.1 Results of MNRU Listening Effort Test

The mean listening effort scores over the five listening levels for each treatment were :-

First Repetition											
	R1	M1	M2	M3	M4	M5	T1	T2	T3	T4	T5
Mean	3.71	3.69	3.69	3.44	2.91	1.40	2.95	2.80	1.27	1.27	0.89
Second Repetition											
Mean	3.71	3.47	3.25	3.16	2.67	1.44	3.02	2.40	1.22	1.13	0.82

Mean scores are based on eleven subjects scores for five listening levels, i.e. 55 scores per treatment. The treatments were placed in rank order and plotted as a column graphs, figs. (5.32a-b).

Column Graphs of MNRU Listening Effort Test Mean Scores

MNRU Listening Effort Test Mean Scores Repetition 1

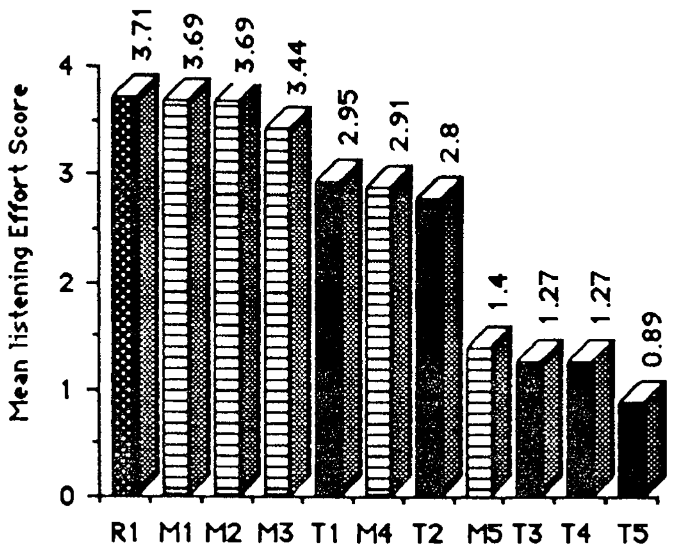


Figure (5.32a)

MNRU Test Mean Scores Repetition 2

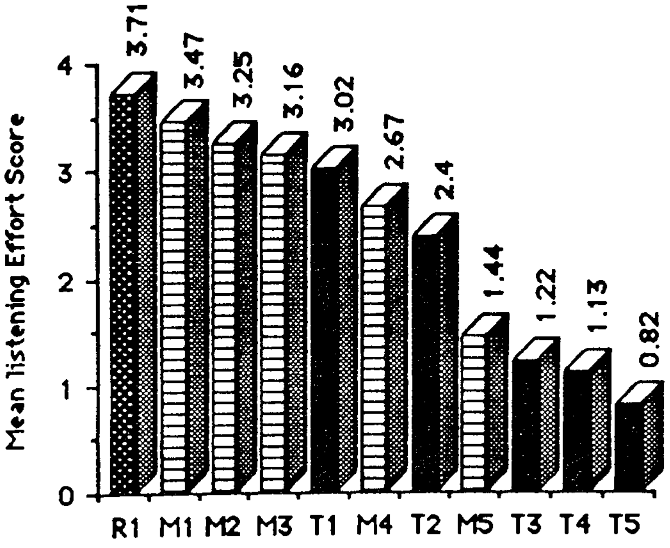
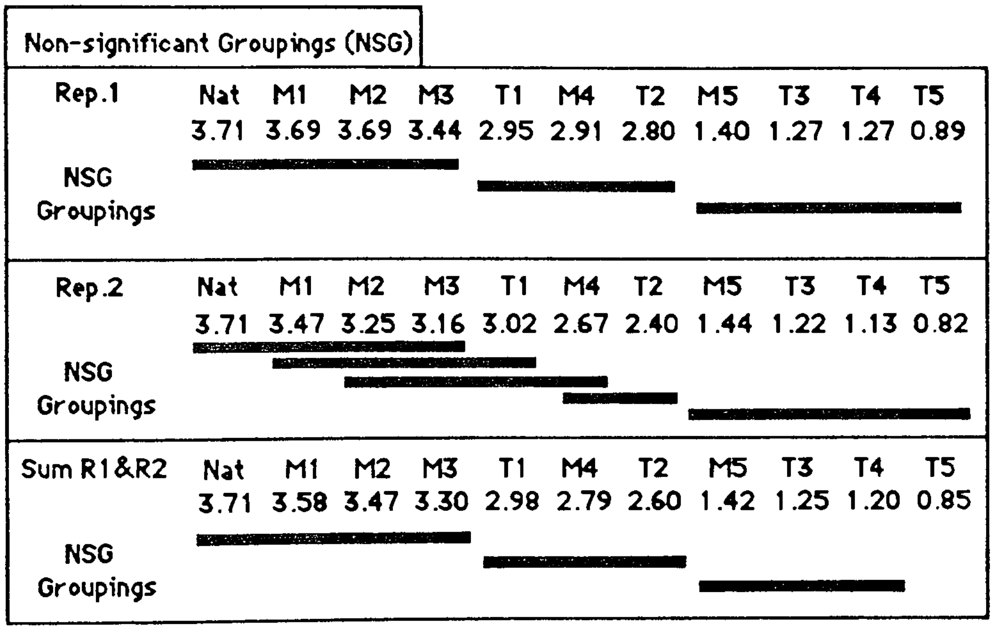


Figure (5.32b)

5.5.9.2 Groupings of Not Significantly Different Treatments

The rank order of the treatments remained the same for the two repetitions of the test. The non-significant grouping of treatments were calculated and are shown in fig. (5.33).

MNRU Listening Effort Test NSG of Treatment Mean Scores



Non-significant Groupings (NSG)

Rep.1	Nat	M1	M2	M3	T1	M4	T2	M5	T3	T4	T5
	3.71	3.69	3.69	3.44	2.95	2.91	2.80	1.40	1.27	1.27	0.89
NSG Groupings											
Rep.2	Nat	M1	M2	M3	T1	M4	T2	M5	T3	T4	T5
	3.71	3.47	3.25	3.16	3.02	2.67	2.40	1.44	1.22	1.13	0.82
NSG Groupings											
Sum R1&R2	Nat	M1	M2	M3	T1	M4	T2	M5	T3	T4	T5
	3.71	3.58	3.47	3.30	2.98	2.79	2.60	1.42	1.25	1.20	0.85
NSG Groupings											

Figure (5.33)

The MNRU treatments M1 M2 and M3 are not significantly different from the natural speech treatment. If we look NSG for the sum of the two repetitions it can be seen that M1, M2, M3, and the natural treatments scored significantly higher than any of the TFW treatments. A complete analysis of variance of the subjective data from the MNRU listening effort tests was carried out. The analysis tests whether the main effects and or the interactions are significantly large.

5.5.9.3 Complete Graeco Latin Square x Layers Analysis of Variance

MNRU Listening Effort Test Rep. 1

Factor	d.f.	SS	MS	F-ratio	Sig
Subjects	10	89.29	8.93	6.309	***
Sequences	10	5.07	0.51	0.358	NS
Treatments	10	683.00	68.30	48.26	***
Lists	10	20.53	2.05	1.450	NS
Error(1)	80	113.22	1.42	5.548	***
Levels	4	10.07	2.52	9.871	***
Interactions					
Subjects x Listening Level	40	15.64	0.39	1.532	*
Sequences x Listening Levels	40	9.13	0.23	0.894	NS
Treatments x Listening Levels	40	17.38	0.43	1.703	**
Lists x Listening Levels	40	8.95	0.22	0.877	NS
Error(2)	320	81.64	0.26		
Total	604	1053.91			

MNRU Listening Effort Test Rep. 2

Factor	d.f.	SS	MS	F-ratio	Sig
Subjects	10	116.45	11.64	8.259	***
Sequences	10	4.34	0.43	0.308	NS
Treatments	10	609.43	60.94	43.22	***
Lists	10	26.92	2.69	1.909	NS
Error(1)	80	112.8	1.41	4.590	***
Levels	4	12.85	3.21	10.45	***
Interactions					
Subjects x Listening Level	40	20.82	0.52	1.695	**
Sequences x Listening Levels	40	10.93	0.27	0.890	NS
Treatments x Listening Levels	40	16.57	0.41	1.348	NS
Lists x Listening Levels	40	8.53	0.21	0.694	NS
Error(2)	320	929	0.31		
Total	604	1037.94			

Chapter 6

Discussion and Conclusions

6 Discussion of Results

6.0 General

British Telecom propose to use listening effort tests with a reference device to provide a standard test for rating the quality of synthesized speech over a telephone connection. In the past telephone engineers have used reference devices for rating the performance of telephone connections. Generally the choice of reference device is governed by the predominant degradation and the type of perceptual errors it produces. None of the reference devices previously used seem particularly suitable for use with synthesized speech. British Telecom Laboratories Martlesham Heath have proposed a new reference device, based on Time Frequency Warping (TFW) for assessing synthesized speech. This study assesses the suitability of TFW modulation for use as reference device. TFW modulation has three degradation parameters, modulation waveform, amplitude and period. It is desirable for a reference device to have one variable parameter as this simplifies test procedure and interpretation of the results. This study determines which TFW parameter should be variable and fix the values of the other two. Having determined the variable parameter, an assessment was made of the suitability of the TFW modulation combined with a listening effort test for assessing the quality of synthesized speech. Assessing the suitability, involved determining the following:-

(i) Perceptual relationship between the TFW modulated, synthesized and natural speech.

(ii) Attributes of speech used by subjects to differentiate between the different types of speech.

(ii) Consistency of rating the synthesized speech in terms of TFW modulation. (TFW modulation equivalent scores)

(iv) Repeatability of listening effort test procedure, i.e. repeatability of the TFW modulated and synthesized speech listening effort scores.

(v) Wether the variance for the mean listening effort scores were significantly different for the different types of speech.

6.1 TFW variable parameter

One of the objectives of this study was to determine which of the time frequency warping (TFW) modulation parameters was to be used as the variable parameter of a proposed reference device. TFW modulation has three possible variable parameters, type of modulation waveform, period and amplitude. The type of modulation waveform was determined by the simple pair comparison test, sec. (5.1). From the choice of square, trianglular and sinewave modulation. Sinewave TFW modulation was chosen because the sinewave modulation had the largest range of perceived difference scores and contained the treatment that was perceived to be the least different from the synthetic treatment. These two properties are important because the TFW modulation has to produce a range of speech quality that is equivalent to the range of speech measured on a listening effort scale. More over, if the two types of speech are perceived to be similar the subjects task of rating the two types of speech in one experiment is made easier.

The next problem was to determine whether the period or amplitude of the TFW modulation was to be the variable parameter. A listening effort test using several settings of TFW modulation period and amplitude was carried out, sec.(5.4). The results were analysed using a multiple regression analysis of the listening effort data. Based on the multiple regression equation a graph was plotted that related mean listening effort scores to various TFW modulation period and amplitude settings, fig.(5.18), sec.(5.4.1.3). The graph allowed the prediction of the effect on listening effort scores of various combinations of TFW modulation period and amplitude settings. The multiple regression model showed that varying TFW amplitude with a fixed period of 150 ms would produce scores that covered the whole listening effort range. Therefore it was decided that the period should be fixed and amplitude used as the variable parameter.

6.2 Attributes of speech used to differentiate between synthesized speech treatments

Another aspect of this study was to identify the perceptual relationship between synthetic and TFW modulated speech. This involved identifying the attributes of speech subjects used to differentiate between the two types of speech. In the simple pair comparison test, subjects rated the "perceived difference" between the TFW modulated and synthetic speech treatments. The use of a non specific criteria, such as perceived difference, meant that the subjects used there own criteria for differentiating between the treatments. The subjects were probably unaware of the particular criteria, or specific attributes of the speech treatments they were using to differentiate between the different types of speech.

The two multidimensional and semantic differential scaling (MDS-SDS) experiments were used to assess particular attributes of synthetic and TFW modulated speech. These experiments determined which attributes were most important in determining the perceived difference. The first MDS-SDS experiment sec.(5.2) measured the effects of simple modifications made to the phoneme parameters of the "British Telecom synthesis system".

The results of the Semantic differential scaling test which measured four attributes of the speech treatments, were used to plot treatment profiles, sec.(5.2.2), fig(5.3). For the natural treatment the measured attributes of naturalness, pleasantness, intelligibility and distinctness all had similar scores. However for the synthetic treatments intelligibility and distinctness attributes were scored higher than the naturalness and pleasantness attributes. This result suggested that the synthetic treatments were lacking naturalness and pleasantness. It should be noted that the quality of the synthesized treatments was not very good compared to the natural. The analysis of variance and least significant difference fig.(5.4 a,b,c,d) showed that the unmodified synthetic treatments were rated higher on all of the semantic differential scales than the modified synthetic treatments.

This implied that the modifications reduced the quality of the British Telecom synthesis system. However it should be noted that the degradation, is not statistically significant. Principal component analysis sec.(5.2.5) of the semantic differential scales data revealed that the first component accounted for 92% of the variance of the data. It can be seen that the synthetic speech treatments had a higher correlation with the first principal component than the natural treatment, but had very low correlation with the second principal component. If we consider the treatment correlations with each principal component as a rating of the treatment. Then a comparison of the rank order of the ratings to the rank order of the treatment ratings on the semantic differential scales can be performed. This comparison shows that the rank order of treatments on the second principal component corresponds closely to the rank order of the treatments on the four semantic differential scales. This suggests that the variance accounted for by the second principal component is due to the variance of the four attributes measured on the semantic differential scales. The first component corresponds to the variance of the differences between the natural treatment and the seven synthesized treatments.

The results of the multidimensional scaling test were used to plot a configuration that models the perceptual relationship of the eight treatments, based on the perceived differences between pairs of treatments sec.(5.2.4). The configuration shows that the perceived differences between the synthesized treatments and natural speech treatment are large in comparison to the perceived differences between the synthesized treatments. The multidimensional scaling configuration fig.(5.6), sec.(5.2.4) shows that the modifications to synthetic treatments S1 & S4 accentuated the perceived difference between the text input and phonetic input treatments. An insight in to what attributes of the speech subjects were basing their opinion of the perceived difference between treatments can be gained from the principal component results. The rank order of the synthetic treatments along the x axis of the configuration fig.(5.6), sec.(5.2.4) is similar to the rank order for the semantic differential scales (second principal component). The y axis corresponds to the perceived differences between the natural treatment and the synthesized treatments. Hence it can be assumed that the dimensions of the configuration correspond to the first two principal components.

6.3 Attributes of speech used to differentiate between synthetic, TFW modulated and natural speech treatments

The objective of the second multidimensional and semantic differential scaling experiment was to, (i) compare specific attributes of TFW modulated and synthetic speech and (i) determine which are used by subjects when determining the perceived difference between the two types of speech. Fig.(5.11), sec(5.3.2) shows the treatment profiles based on the five attributes measured in the semantic differential scaling parts of the experiment. The general quality of the synthetic and, hence, the TFW modulated speech was fairly poor compared to natural speech. The TFW modulation period and amplitude parameter settings were chosen so that the quality of the modulated speech was similar to that of the synthetic speech. The low quality of the speech is reflected in the treatment profiles by the low scores on the semantic differential scales for the TFW modulated and the synthesized speech. The natural treatment profile was distinguished from the other treatments because the scores on the semantic differential scales were much greater. Moreover the measured attributes were perceived to be present in natural speech in approximately equal amounts. Whereas, the synthetic treatments the intelligibility and distinctness were relatively higher than the naturalness and pleasantness attributes of the speech. With the listening effort (lack of) attribute score being somewhere between the naturalness and intelligibility scores. These differences were not present to such an extent in the TFW modulated treatments, however it was noted that the distinctness scores decreased quite dramatically with respect to the other attributes when the values of the modulation parameter were increased. Tables.(5.9a-e), sec.(5.3.3) show the non significant (not significantly different) groupings of treatment mean scores for each semantic differential scale. It can be seen that for all five of the attributes the treatments are in large overlapping non significant groups, this implies that any conclusions about such treatments are not very accurate. However, in general, it can be stated that the synthetic treatments were rated higher on the intelligibility and listening effort (lack of listening effort) scales than the TFW modulated speech. In contrast the TFW modulated speech treatments were rated higher on the naturalness and pleasantness scales.

6.3.1 Perceptual relationship of TFW modulated, synthesized and natural speech

The MDS configuration calculated from the results of the multidimensional scaling part of the experiment is shown in fig.(5.12), sec.(5.3.4). The MDS configuration maps the perceptual relationship of the speech treatments based on the perceived differences between pairs of treatments. The synthetic treatments are grouped together away from the natural and TFW modulated speech. This shows that the subjects perceived the synthesized speech to be different from the natural and TFW modulated speech. In order to interpret the MDS configuration, a second configuration was calculated using the semantic differential scaling (SDS) data. The results of the principal component analysis of the SDS data were used to calculate another configuration sec. (5.3.6.1). The two configurations were compared by plotting them on the same axes figure (5.14). The similarity of the configurations is surprising considering the fact they are based on two different types of data, such as SDS and MDS. The similarity of the configurations suggests that subjects were basing their opinions of the perceived difference between treatments on the attributes measured in the SDS tests, i.e. naturalness, pleasantness, intelligibility, distinctness and listening effort. Figure (5.16) shows the semantic differential scales plotted on multidimensional scaling configuration. It can be seen that the main difference between the TFW modulated and synthetic speech treatments is based on the naturalness and pleasantness of the speech. In comparison with the natural speech the intelligibility of the TFW modulated and synthetic speech treatments is quite low.

6.4 Relationship between TFW modulation variable parameter and listening effort scores

The objective of this study was to assess the suitability of Time Frequency Warping modulation as a reference device for rating speech synthesis systems. A reference device degrades natural speech by different amounts corresponding to the setting of the reference device variable parameter. For TFW modulation the amplitude of the modulation was selected as the variable parameter, sec. (5.1). Fig.(5.18), sec. (5.4.1.3) plots the mathematical model relating listening effort score to the modulation amplitude and period settings. Fig.(5.18) shows that increasing the amplitude of the modulation decreased the listening effort score. For TFW to be a viable reference device the relationship between the TFW modulation parameter setting and listening effort score must be preserved. If we look at the rank order of TFW treatments mean listening effort scores, for each of the four comparison listening effort tests, figs.(5.19, 5.22, 5.25, 5.28). It can be seen that the relationship between TFW modulation amplitude settings and listening effort scores is essentially preserved, i.e. increasing the TFW modulation variable parameter decreases listening effort scores. However this relationship is violated for the first repetition of comparison listening effort test three. Where T3 (TFW amplitude 0.75 KHz) ranked higher than T2 (TFW amplitude 0.5 KHz). However it should be noted that the mean scores for T2 & T3 were not significantly different, fig.(5.26), sec.(5.5.3.2).

6.5 Repeatability of TFW modulation and listening effort test procedure

For a listening effort test, combined with TFW modulation, to become a standard procedure for rating synthetic speech quality the test must be repeatable. That is the procedure should produce similar listening effort scores for the same setting of the reference device for tests carried out at different times and places. In this study eight comparison listening effort tests were carried out (4 x 2 repetitions). If we consider three settings of TFW modulation, i.e. 500 Hz, 1500 Hz, and 2500 Hz, which were used as treatments in all eight comparison listening effort tests, the listening effort scores for each setting over the eight tests can be compared. The listening effort scores compared are not the subject's mean scores for the TFW modulation settings, but the listening effort scores predicted from the regression line for the TFW settings. These were termed predicted listening effort scores (P-LE) sec.(5.5.5). The regression line scores (P-LE) were compared because the regression line is based on seven listening effort mean scores, which would be more accurate than comparing single mean scores. Also it is the regression line that is used to calculate TFW equivalent scores which are used to rate synthetic speech, sec. (4.5.1). The predicted listening effort scores of the three TFW modulation settings for the eight comparison listening effort tests were calculated, sec.(5.5.5). The mean and range of the P-LE scores were calculated and compared with the regression line confidence limits of three comparison listening effort tests. The results showed that for TFW modulation settings 1500 Hz and 2500 Hz the range of P-LE scores over the eight tests fall within the regression line confidence limits of the individual tests at those TFW settings. For TFW setting 500 Hz the P - LE range of scores is greater than the confidence limits of the individual test shown in the table. The largest deviation is 13 points on the transformed listening effort scale. This is approximately 0.2 on the untransformed listening effort scale. If this value is compared to the values for the difference between treatment mean scores in the non-significant grouping tables (5.20), sec.(5.5.1.2), it is seen that a difference of 0.2 between treatments is not significant. This means that the P-LE scores which fall outside the individual tests regression line confidence limits are not significantly different from P - LE scores that have the same scores as the confidence limits. The comparison of P-LE scores over the eight comparison listening effort tests shows that the listening effort TFW modulation procedure produces repeatable results.

6.6 Consistency of listening effort scores for synthesized speech

The consistency of listening effort scores for a particular synthesis system was looked at in section (5.5.6). The synthesis system "INFOVOX 850905 Preliminary British" was used in three of the comparison listening tests. The transformed listening effort score over the six tests (3 x 2 reps.) were similar with a range from 452 - to - 472 on the transformed listening effort scale. This corresponds to a difference of 0.23 on the untransformed listening effort scale. If we compare the range of listening effort scores for "INFOVOX 850905 "Preliminary British" over the six tests, with the confidence limits of the listening effort scores for "INFOVOX 850905 Preliminary British" in each test. In most cases the range of listening effort mean scores for the synthesis system fall within the confidence limits of the listening effort mean scores of the individual tests. However the mean score of the synthesizer in comparison listening effort test 4 rep.2 (472), is just above the upper confidence limit for the mean score in comparison listening effort test 2 rep.1 (466), the difference is negligible. This suggests that the differences in the listening effort mean scores for the "INFOVOX 850905 Preliminary British" system for different tests is comparable to the expected differences with in a test. The TFW modulation equivalent settings, TFWe, were calculated for the "INFOVOX 850905 Preliminary British" system for the six comparison listening effort tests, sec.(5.5.6). TFWe is the setting of the TFW modulation parameter that would produce the same mean listening effort score as the "INFOVOX 850905 Preliminary British" system. For TFW modulation and listening effort test to be a viable standard test procedure for rating synthesis systems, it must produce consistent equivalent setting of the degradation variable parameter, for a synthesizer, over a number of repeated tests. The TFWe setting for the "INFOVOX 850905 Preliminary British" system over the six tests were calculated sec.(5.5.6). The results showed that the range of TFWe settings over the six tests were comparable with confidence limits of a TFWe setting with in a single test.

6.7 Rating six synthetic speech systems in terms TFW variable parameter.

The TFW modulation equivalent settings TFWe were calculated for six of the synthesis systems used in the comparison listening effort tests, sec.(5.5.7). The speech over a telephone connection must have a listening effort score over 2.5 to be considered for use in the public network. A listening effort score of 2.5 has a TFW equivalent of approximately 1 KHz. Of the six synthesizers tested only the DEC Talk system produced speech that would considered for use on the public network.

6.8 Comparison of subject variance for TFW modulated, synthesized and natural speech

Individual subject's listening effort scores for different types of speech are dependant on the individuals abilities and experiences i.e. familiarity with that particular type of speech. For example an airline pilot easily understands the poor quality speech transmitted from the tower, whereas most people would require a lot of effort. Although this is an extreme example, it does highlight the effects poor quality speech has on individual subject's. The effects of differing speech quality and the differences between subjects can be seen as differences in the variance of the mean score for different types of speech. If subjects are rating high quality natural speech you expect the variance of the mean score to be small; similarly a low value could be expected for very low quality speech. For each comparison listening effort test the variance of the treatment means (11 subjects over 5 listening levels) were calculated. Bartlet's test app. (A.2.5) was used to determine the homogeneity of the variances, sec.(5.5.71). If the variance for a particular treatment mean is significantly high compared to the other mean variances it indicates that the group of subjects are having trouble rating the speech consistently. The results show that for each comparison listening effort test, the variances of the treatment means were homogeneous, i.e. not significantly different. However, it was noted that in general the variances of the synthesized speech treatments were higher than the other treatments. Indicating that a group of subjects is more consistent rating the listening effort of TFW modulated and natural speech. The multidimensional scaling tests required subjects to rate the perceived difference between different types of speech, this was used to plot a MDS configuration. The analysis of the MDS data only used subjects results if the stress app.() was low enough. Stress measures the individual subjects ability to consistently rate perceived difference. The results showed that there were differences between individual configurations. To compare the configurations, standard configurations were calculated sec.(5.3.5). The variances of the mean differences between the three types of speech were compared. It was found that the variances were not significantly different, implying that a group of subjects could, rate in a particular test, the differences between, natural & synthetic speech, TFW modulated & natural speech, and synthetic & TFW modulated with the same degree of accuracy.

6.9 Conclusions and Further work

6.9.1 Attributes of Speech used to Determine the Perceived Differences Between different Types of Speech

The results of the second multidimensional and semantic differential scaling experiment produced a configuration, sec.(5.3.6) fig.(5.15), based on the MDS perceived difference data and the semantic differential scales data. The configuration represents a two dimensional "cognitive space" that maps the perceived relationship between TFW modulated, synthesized and natural speech. The dimensions of the configuration were determined using the semantic differential scaling data, two axes were fitted intelligibility and naturalness of the speech. It was concluded that subjects differentiated between the different types of speech in terms of the intelligibility and naturalness of the speech. It was found that the attributes, pleasantness and distinctness were synonymous with the naturalness and intelligibility attributes respectively. The MDS configuration sec.(5.3.6), fig.(5.15) demonstrates there is a degree of independence between the perception of the naturalness and the intelligibility of speech. This can be explained by the fact that speech can be intelligible but it does not have to sound natural. For example if every word is over-articulated and spoken at a very slow rate, with long pauses between words, then the speech will be intelligible, but at the same time sounding totally unnatural.

The degree of independence between the naturalness and intelligibility of speech was reported in a paper "Subjective assessment of automatic voice answering machines" ref. (6). The paper presents results which demonstrate how the profiles of suitable attributes illustrate important differences in the perceived characteristics of five voices. It was found that a feature common to the synthesized voices is that the listening effort scores are appreciably higher than the naturalness scores. Indicating that a lack of listening effort (more intelligible) can be partially separated from the naturalness of the speech. Other studies have reported that the naturalness and intelligibility of speech can be partiality separated, Voiers, 1980) ref.(39) and Pisoni ref.(26).

6.9.2 Similarity of TFW Modulated and Synthetic Speech

The aim of this study is to assess the suitability of Time Frequency Warping modulation for use as a reference degradation device for assessing text-to-speech synthesizers, using listening effort comparison tests. For subjects who take part in such tests their task is made easier if the reference device produces speech that is similar to the speech under test. Hence part of this study was to: (i) determine the perceived similarity between TFW modulated and synthesized speech, (ii) identify the attributes of the speech on which the perceived differences are based. The MDS configuration fig.(5.12), sec.(5.3.4), shows the perceptual relationship between TFW modulated, synthesized and natural speech.. An interpretation of the MDS configuration fig.(5.15), sec.(5.3.6), shows that the perceived differences between the three types of speech are mainly based on two speech attributes, naturalness and intelligibility. The TFW modulated speech was perceived to be more natural but less intelligible than the synthesized speech. Other studies Voiers, 1980) ref.(39) and Pisoni ref.(26) concluded that the intelligibility of speech is dependant mainly on the segmental intelligibility and naturalness of speech is dependant on the suprasegmental prosodic qualities. Therefore we may be able to assume that it is the suprasegmental properties of the synthetic speech that are degraded compared to natural and TFW modulated speech. It may also be assumed that the TFW modulation of natural speech mainly degrades the segmental properties of natural speech, i.e. the acoustic-phonetic cues. It should be pointed out that the quality of the synthesized speech used in the MDS - SDS test was quite poor compared to natural speech. This was shown in the first comparison listening effort test sec.(5.5.1.1) which used the same samples of synthesized speech. If higher quality synthetic speech had been used, the MDS - SDS test, the TFW modulated and synthesized speech may have been perceived as more similar. However the MDS - SDS test did demonstrate that the test subjects were able to consistently rate the perceived differences between the different speech samples.

6.9.3 Consistency of Subjective Scores

If subjects are having to construct different "perceptual spaces" and use different perceptual strategies to understand the individual speech systems, then the "cognitive load", ref.(26) associated with the perceptual processing of the different speech systems could have an effect on the accuracy of subjective speech quality tests. The increased "cognitive loading" associated with the perception of synthesized speech could accentuate subject variance, i.e. the variance due to individual subjects within a test group. Increased variance means that the subject's scores within a group of subjects are less consistent. In this study we compared the variance due to test subjects for the TFW modulated, synthetic and natural speech treatments used in the comparison listening effort tests, sec.(5.5.71). It was found, that although the subject variance of the listening effort scores for the synthesized speech treatments were, in general, larger than the TFW modulated and natural speech treatments, the difference was not significant.

6.9.4 Combined MDS and SDS Test as a Diagnostic Tool

A combined MDS & SDS test would be useful for engineers involved in the development of speech devices. For example, if an engineer was trying to improve the quality of speech produced by a synthesizer and modifications were made to the synthesizer, then a MDS configuration would show whether the modifications had made any perceivable differences to the output speech compared to the unmodified synthesizer. More specifically, if modifications were made to the suprasegmental parameters of the synthesizer to improve the naturalness of the speech, then a MDS configuration can show the effects on the naturalness and intelligibility of the speech, and whether the difference was perceived by the test subjects. If several different sets of modifications were made, the perceptual relationship between the different modifications can be mapped.

The results of the second MDS and SDS experiment showed that subjects were rating the perceived differences between synthetic,TFW modulated and natural speech using a two dimensional perceptual space based on the naturalness/pleasantness and the intelligibility/distinctness of the speech. Whereas in the first MDS and SDS test, the speech treatments were very similar. the differences were rated mainly on a single dimension. The MDS - SDS tests are very sensitive to small changes in the speech quality as demonstrated in the first MDS and SDS test, in which they were able to differentiate between the phonetic and text input speech treatments.

6.9.5 Time Frequency Warping Modulation for Assessing the Quality of Synthetic Speech

Several factors have to be taken into consideration when considering the suitability of TFW modulation for use as a reference device for assessing synthesized speech. Firstly whether the variable parameter (amplitude) produced a sufficient range of degradation in natural speech to cover the whole listening effort scale. Secondly whether the variable parameter (degradation) settings could produce consistent listening effort scores. Thirdly whether synthetic, TFW modulated and natural speech could be rated consistently in terms of listening effort by a group of test subjects. This study has shown that TFW modulation could produce a sufficient range of degradation in natural speech to cover the whole listening effort scale. The results of the comparison listening effort tests showed that the TFW modulated speech treatments were rated consistently in each test, i.e. the relationship between TFW modulation setting and listening effort scores were preserved. Also that the range of scores for individual TFW modulated and synthetic speech treatments between tests were similar to the expected differences within a single test. The comparison listening effort tests showed that the listening effort tests combined with TFW modulation could consistently rate the quality of synthesized speech over repeated tests. Groups of subjects were able to rate the different types of speech consistently. That is the subject variances for the different types of speech were not significantly different.

6.9.6 Suitability of TFW modulation and listening effort test as standard procedure for rating the quality of synthesized speech

The main purpose of this study was to determine the suitability of Time Frequency Warping for used as reference degradation for rating the quality of synthesized speech. The results of this study show the perceptual relationship between three types of speech, TFW modulated, synthetic and natural in a two dimensional cognitive space. The dimensions of the cognitive space are the intelligibility and naturalness of speech. The positions of the different types of speech depends on the intelligibility and naturalness of the speech as perceived by the test subjects. It is desirable for a reference degradation to to produce speech that is perceived to be similar to the speech it is rating. Although the TFW modulated speech was not perceived to be similar to the synthesized speech, it is not so different, that groups of subjects have problems in rating the two types of speech on the same scale. The homogeneity of the variances of the mean scores in the comparison listening effort tests, demonstrate that groups of subjects are able to rate TFW modulated and synthetic speech with similar degrees of accuracy. The comparison listening effort tests showed that a TFW modulated degradation and listening effort test could be used as a standard procedure for rating the quality of synthesized speech. The comparison listening effort tests showed that the procedure could consistently rate (i.e. tests were repeatable) the quality of synthesized speech in terms of the reference degradation.

6.9.7 Further Work

Further listening effort tests should be conducted using higher quality synthesized speech. The listening effort tests, should also compare the degradation in speech quality in noise of speech synthesis systems and their TFW equivalent speech treatment. MDS & SDS tests using various examples of high quality synthetic speech and natural speech should be conducted to determine whether higher quality synthetic speech is perceived similar to TFW modulated and natural speech.

Appendix

Analysis Theory, Computer Procedures and Subject Data

Appendix

A.1 English Phonemes

Table (A.1) shows 49 English phonemes, the phonemes are coded using two codes, computer coded phonemes (CPC) and the international phonetic alphabet (IPA). (EE) means English equivalent.

CPC	EE	IPA	CPC	EE	IPA	CPC	EE	IPA	CPC	EE	IPA
Vowel Sounds											
AY	pale	eɪ	EH	get	ɛ	UW	you	iu-	O	cot	a-
AE	Black	æ	ER	perk	ɜ	/U	put	u-	UX	coot	u
AA	car	ɑ	IY	site	aɪ	UH	wood	u-	OY	voice	ɔi
AI	fair	–	IX	sit	ɪ	AH	up	v-	AW	now	ɑʊ
EE	meet	i	IH	sit!	-	OW	coat	ou	AO	door	ɔ
									OH	won	
Consonants											
B	bat	b	J	jab	dʒ	P	pat	p	W	wag	w
D	dab	d	K	cat	k	R	rat	r	Y	yap	j
F	fat	f	L	lag	l	S	sat	s	Z	zap	z
G	gap	g	M	mat	m	T	tap	t			
/H	hat	h	N	nap	n	V	vat	v			
CH	chair	tʃ	TH	thick	ð	DR	dragon	–	TR	track	
DH	this	θ	ZH	azure	ʒ	DUX	duke	–			
SH	share	ʃ	CT	fact	-	NX	sing	ŋ			

* short 'i'

** long 'i'

*** quiet 'u'

A.2 Analysis Theory

A.2.1 Multidimensional Scaling Theory

A full discussion of the principles of multidimensional scaling are best described by Kruskal & Wish ref (29). Below is a summary of multidimensional scaling theory used in this study. Before a discussion of the theory of multidimensional scaling several terms must be defined.

OBJECT - the thing or event to be investigated, in the context of this work they are short samples of speech. These speech samples are also referred to as speech treatments.

PROXIMITY - This is a measure of the distance between objects. In this study proximity refers to subjects perceived difference between two speech treatments.

DATA MATRIX - Multidimensional scaling operates on the proximities associated with pairs of objects. Proximities amongst N objects can be represented as a N * N matrix, where the entry in row i and column j, mij is the proximity of object i to object j. It generally can be assumed that mij is equal to mji so that the data matrix is symmetrical. Hence the data matrix can be reduced to a trianglular matrix of (N(N-1)/2) proximities.

A data matrix for 5 objects would be;

m12	m13	m14	m15	m - proximity distance
	m23	m24	m25	Numbers refer to objects
		m34	m35	
			m45	

CONFIGURATION - the output of a multidimensional scaling analysis is a table of coordinates which locate each object in multidimensional Euclidean space. A configuration represents the multidimensional Euclidean space. The distance between each object in the configuration corresponds to or is a function of the proximity associated with the two objects. The distances can be arranged as a matrix similar to the data matrix. ie distance matrix for 5 objects

d12	d13	d14	d15	d - distance
	d23	d24	d25	Numbers refer to objects
		d34	d35	
			d45	

OBJECTIVE FUNCTION - The objective function is a trial proximity data transfer function.

STRESS - Stress is a measure of "goodness of fit" of the distances calculated from the configuration and the proximity data.

The central concept of multidimensional scaling is that the distance dij between the points i and j on the configuration correspond or are a function of the perceived proximity mij between the objects i and j. Multidimensional scaling can be divided into to two broad categories metric and non-metric. Mapping the proximities data matrix into distances on the configuration generally requires a transformation of the proximities. If the transformation function is linear then the scaling is metric. ie;

$$b\ (mij) = dij \quad - (i)$$

If the function is monotonic, ie

$$f^{`}\ (mij) = dij \quad - (ii)$$

Where $f^{`}$ is a function of a curve or series of curves, or just a rising function, then the scaling is non-metric.

Defining an Objective Function.

It is believed that a clear separation of definition of the objective function and the computer procedure has several important advantages, and historically there has been a strong trend in this direction. This provides a uniform approach to the different types of multidimensional scaling. For any set of data and for any configuration, the objective function yields a single number which shows how well the transformed data fits the configuration. The basic concept takes the form;

$$f^{`}\ (mij) = dij$$

Where f is a transformation function of some specified type. One natural objective function can be formed as follows: suppose we have some function f, of the specified type. The discrepancy between f(mij) and dij is then,

$$f^{`}\ (mij)\text{-}dij$$

These discrepancies are vertical line segments for a function f which is increasing. The size of the discrepancy is measured by taking the square, since positive and negative discrepancies are equally undesirable. The sum of square of the discrepancies for all proximities yields the formula

$$\sum_{i}\sum_{j}\left[(f^{`}\ (mij)) - dij \right]^{2}$$

Next we divide by a scale factor in order to measure the squared discrepancies relative to a sensible measuring stick. The most commonly used is,

$$\sum_{i}\sum_{j}(dij)^2$$

Finally for minor reasons, take the square root of the result. Thus an "objective function" is obtained which is called "f - stress". The formula for f - stress is,

$$\sqrt{\frac{\sum_{i}\sum_{j}\left[f(mij) - dij\right]^2}{\text{Scale factor}}}$$

The larger the f - stress, the worse the configuration and the function f fit the proximity data. The measure can be defined as;

STRESS = min f - stress over all f

which simply says that the best possible f is used for this configuration in measuring how well the configuration matches the data. (Different functions are best for different configurations)

Computional Procedures

The computional procedure developed by Shephard and Kruskal ref. (29), was to find the best configuration and function by minimizing f- stress over all possible functions. For a given proximity data matrix a analysis is one of the simplest forms of factor analysis. Factor analysis postulates a model of the form

$$x = a\,F + e \qquad (i)$$

x - data

F - Factor scores

a - Factor loadings

i - observation index

j - data variable index

k - factor index

e - model error

Geometrically, Eq. (i) can be interpreted as projecting the data (from j space) onto a lower k-dimensional factor space, in which the factor loadings form the basis of the factor space and the factor scores are the coordinates which locate the data in the factor space. Ideally the elements of the factor loadings (aij) are either very large or very small, so that each data variable is associated with a minimum number of factors. This can be achieved by rotation of the factor loadings to form a new but equivalent basis.

A.2.2 Principal Component Analysis

In the principal component method of analysis, ref Morrison (20) a sample covariance matrix is calculated for all variables in the analysis. The matrix is normalized so that the determinant equals 1.0 and the eigenvalues and eigenvectors extracted. Each eigenvalue greater than 1.0 indicates a factor, the square root of the eigenvalue multiplied by the associated eigenvector is the factor loading. The variance accounted for by each factor is equal to the eigenvalue associated with that factor. To minimize the number of dimensions of the factor space, eigenvectors with eigenvalues less than 1.0 are interpreted as aspects of the data that are not significant. Factor loadings are rotated, and factors and loading elements within each factor are sorted in order of decreasing significance.

A.2.3 Multiple Regression

Multiple regression, Morrison ref.() can be used to analyse data in which the dependent variable is affected by several controlled variables. For example a linear regression equation of three control variables will take the form;

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3$$

Given n sets of observations

$$(y_1, x_{11}, x_{21}, x_{31}) \ldots\ldots\ldots\ldots, (y_n, x_{1n}, x_{2n}, x_{3n})$$

the least squares estimates of can be obtained in a similar manner to linear regression. The sum of the squared deviations of the observed values of y from the predicted values is given by

$$S = \sum(y_i - a_0 + a_1x_{1i} + a_2x_{2i} + a_3x_{3i})^2$$

This quantity can be minimized by setting $\frac{\partial S}{\partial a_0}, \frac{\partial S}{\partial a_1}, \frac{\partial S}{\partial a_2}$ and $\frac{\partial S}{\partial a_3}$ equal to zero, to obtain the four simultaneous equations in $\hat{a}_0$, $\hat{a}_1$, $\hat{a}_2$ and $\hat{a}_3$.

$$\hat{a}_0 n + x + \hat{a}_1 \sum x_{1i} + \hat{a}_2 \sum x_{2i} + \hat{a}_3 \sum x_{3i} = \sum y_i$$

$$\hat{a}_0 \sum x_{1i} + \hat{a}_1 \sum x_{1i}^2 + \hat{a}_2 \sum x_{2i} x_{1i} + \hat{a}_3 \sum x_{3i} x_{1i} = \sum y_i x_{1i}$$

$$\hat{a}_0 \sum x_{2i} + \hat{a}_1 \sum x_{1i} x_{2i} + \hat{a}_2 \sum x_{2i}^2 + \hat{a}_3 \sum x_{3i} x_{2i} = \sum y_i x_{2i}$$

$$\hat{a}_0 \sum x_{3i} + \hat{a}_1 \sum x_{1i} x_{3i} + \hat{a}_2 \sum x_{2i} x_{3i} + \hat{a}_3 \sum x_{3i}^2 = \sum y_i x_{3i}$$

These four equations, called the normal equations , can be solved to give the least squares estimates of

$$a_0, a_1, a_2 \text{ and } a_3.$$

The equations were solved using a computer program.

A.2.4 Complete Analysis of Variance Theory

The analysis is based on a 11 x 11 graeco-latin square experimental design. All of the listening effort tests carried out in this study were based on randomized 11 x 11 graeco-latin square designs.

The following sub totals were calculated.

$S_{in} = \sum_{j=1}^{11} Y_{ijn}$ Sum of the i th subject`s scores over all columns at the n th level

$C_{jn} = \sum_{i=1}^{11} Y_{ijn}$ Sum of all the subjects scores in the j th coloumn at the n th level.

$U_{rn} = \sum_{i=1}^{11} Y_{irn}$ Sum of all the subjects scores using the r th list at the n th level.

$L_{n} = \sum_{n=1}^{11} S_{in}$ Sum of all the subjects scores over all columns at the n th level

$S_{i} = \sum_{n=1}^{5} S_{in}$ Sum of the i th subject's scores over all columns over all levels

$C_{j} = \sum_{n=1}^{5} C_{jn}$ Sum of all subjects scores in the j th column over all level

$U_{r} = \sum_{n=1}^{5} U_{rn}$ Sumof all subjects scores using the r th list over all level

$G = \sum_{i} \sum_{j} \sum_{n} Y_{ijn}$ Sum of all scores, and let $Gq = G^2/605$

These sub totals are used in the analysis of variance table below.

Factor	D.F.	Sum of Squares
Listening levels	4	$sq1 = (\sum_{n=1}^{5} L_n^2)/121-Gq$
Subjects	10	$sq\,2 = (\sum_{i=1}^{20} S_i^2)/\,55-Gq$
Interaction (Subjects & levels)	40	$sq\,3 = (\sum_{n=1}^{5}\sum_{i=1}^{20} s_i^2)/11-Gq-sq1-sq2$
Columns	10	$sq\,4 = (\sum_{i=1}^{20} c_j^2)/55-Gq$
Interaction (Columns & levels)	40	$sq\,5 = (\sum_{n=1}^{5}\sum_{j=1}^{20} c_{jn}^2)/11-Gq-sq1-sq4$
Treatments	10	$sq\,6 = (\sum_{p=1}^{20} K_p^2)/55-Gq$
Interaction (Treatments & levels)	40	$sq\,7 = (\sum_{n=1}^{5}\sum_{p=1}^{20} K_{pn}^2)/11-Gq-sq1-sq6$
Lists	10	$sq\,8 = (\sum_{r=1}^{20} U_r^2)/55-Gq$
Interaction (Lists & Levels)	40	$sq\,9 = (\sum_{n=1}^{5}\sum_{r=1}^{20} U_{rn}^2)/11-Gq-sq1-sq8$

Lists	10	$sq\,8 = (\sum_{r=1}^{20} U_r^2)/55-Gq$
Interaction (Lists & Levels)	40	$sq\,9 = (\sum_{n=1}^{5}\sum_{r=1}^{20} U_{rn}^2)/11-Gq-sq1-sq8$

Residual (d.f.400)

$$sq10 = sq11-sq1-sq2-sq3-sq4-sq5-sq6-sq7-sq8-sq9$$

Total	604	$sq\,11 = \sum_{n=1}^{5}\sum_{i=1}^{20}\sum_{j=1}^{20}(Y_{ijn})^2-Gq$

"

D.F. - "degrees of freedom".

The "mean square" is calculated for each sum of squares except for the residual and total sum of squares by dividing by the appropriate degrees of freedom . For each mean square the F - ratio is calculated by dividing by the residual mean square. The significance of each factor is determined by comparing the factor's F - ratio with the values in a standard table. "Significant " in this context is a technical term meaning that the magnitude of the observed effect is such that there is a probability of less than 0.05 of obtaining an equal or greater observed value by chance. "Highly significant" is similarly defined but with a probability of less than 0.01.

A.2.5 Bartlett`s Test Homogeneity of Variances

The test was used for checking the assumption of a constant (variance) in analysis of variance or regression (homoscedescity). The Bartlett`s test, tests the hypothesis var1 = var2 ..= vark versus var`s not equal. In practice we test the sample variances. In this study all the samples are the same size n. The standard test is known as Bartlett`s test , having been developed by M.S.Bartlett . The formula for determining the homogeneity of variances for samples of equal size is shown below;

$$K_{k-1} = \left[k(n-1)\,\mathrm{Log}_e\left(\sum_{i=1}^{k}\sigma^2/k\right) - (n-1)\sum_{i=1}^{k}\mathrm{Log}_e\,\sigma^2\right]/C \qquad (1)$$

$$K^2_{n-1} = \frac{k(n-1)}{C}\,\mathrm{Log}_e\left[\,\mathrm{AM}\,(s_i^2)/\mathrm{GM}\,(s_i^2)\,\right] \qquad (2)$$

Equation (2) shows that K is basically a comparison of the arithmetic and geometic means (AM, GM) of the variances, where GM(X1...Xk) = (X1..Xk)exp1/k. This measure depends on the fact that, the greater the relative variability in a collection of positive quantities, the greater the ratio of these two averages (means). Hence the greater the relative variation amongst the sample variances the larger value of K. Thus to test the null hypothesis of " All the sample variances are not significantly different "with the alternative hypothesis" the sample variances are "significantly different". We use a one tailed test to determine whether the value of K lies in a region above a selected significance level Bartlett`s test uses the chi - squared distribution. Hence if the value of K was greater than the value of selected significance level at the same number of degrees of freedom, the null hypothesis would be rejected.

A.3 Computer Procedures

A.3.1 MDS Computer Procedure

The experiments carried out in this study used only 8 or less objects which meant that only two dimensions could be justified in fitting a configuration to the data. The program used an iterative process to adjust the coordinates of the points of the configuration using partial derivatives. The iteration process stops when the rms of the partial derivatives of all the coordinates fell below a set value, in this case 0.1, the stress was then calculated. The program does not attempt to find a function to relate the proximity data and the configuration, because the low stress values from the experimental data indicate that the two dimensional configuration corresponds to the experimental proximity data. The stress of a subjects proximity data and the configuration were used as a measure of the subjects ability to give relatively consistent opinions throughout the experiment.

A.3.2 Principal Component Analysis Computional Procedure

A correlation matrix **R** of all the semantic differential scales is calculated using the computer program "EstCorr". The resultant correlation matrix is used as the input data matrix for the program "compute P.C." . The program "compute P.C." uses an iterative process based on Eigenvectors and Eigenvalues to determine the principal components of the input correlation matrix **R**. The iterative process starts by estimating the eigenvector for the correlation matrix R, using a unit column vector. **R** is squared and multiplied by the unit column vector. The resulting column vector is the sum of the columns of **R** squared and is standardized by dividing by the absolute value of the maximum value of **R** squared. The rms. error of the first and second column vectors is calculated. If the rms. error is below 0.001 the iteration stops, if not the second iteration starts by squaring the squared correlation matrix i.e. **R** to the 4th.and multiplied by the standardized column vector. The iterative process is repeated until the rms. error falls below 0.001. The iterative process uses the fact that the true eigenvector of **R** is the same for **R** and **R** squared. The resultant standardized vector **C** from the iteration process is normalized to give column vector **N** and the eigenvalue or characteristic root of the correlation matrix **R** is calculated. The principal component matrix is calculated by multiplying the standardized vector **C**, normalized vector **N** and the eigenvalue together to give **P1** the first principal component matrix. The second principal component matrix is calculated by subtracting **P1 from R** . The residual matrix is used as the input matrix to the iterative process and a similar process to above is carried out. The third and fourth principal components are extracted in a similar manner.

A.3.3 Multiple Regression Computer Procedure

The means and standard deviations of the observed values are calculated for each control variable. The matrix of the coefficients is set-up and the vector of the coefficients is extracted. The predicted values of y are calculated for the coefficient vector and the respective data values. The sum of the predicted y values is calculated . Sum of squares of the predicted y values and the sum of squares of the observed y values were calculated. The sum of squares of the predicted and observed y values was calculated. The variance of the deviations (error), plus the variance of the observed y values were calculated. If the variance of the deviations is less than the variance of the observed y values the multiple correlation coefficient can be calculated.

A.4 Subjects Scores

A.4.1 Simple Comparison Test Subject data

Subjects	S1	S2	S3	S4	T1	T2	T3	T4	R1	R2	R3	R4
1	4.2	5.6	6.3	6.9	7.0	5.2	4.1	5.9	4.2	73	7.7	7.2
2	1.6	4.1	6.7	5.5	7.6	5.5	3.9	6.1	3.3	8.1	6.8	8.3
3	3.0	3.0	3.4	4.4	2.4	2.8	3.0	2.2	5.8	7.7	6.0	7.8
4	2.3	7.2	8.3	8.5	7.1	5.2	7.8	7.3	4.3	8.5	7.5	6.1
5	5.2	5.1	6.3	6.2	6.7	4.6	5.0	5.6	4.6	2.5	4.3	6.4
6	5.4	7.1	5.9	8.3	6.5	4.9	7.5	5.8	5.2	6.5	8.6	6.1
7	2.3	2.5	3.6	1.7	1.8	2.8	0	2.7	3.6	4.0	1.8	3.5

A.4.2 First MDS-SDS Test Multidimensional Scaling and Semantic Differential Scaling Data

A.4.2.1 First MDS-SDS Test Semantic Differential Scaling- Data

Subject 1

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	4.9	2.1	4.3	2.7	2.0	4.8	2.9	9.9
Distinctness	4.1	3.3	4.3	2.8	1.4	6.5	3.4	9.9
Naturalness	3.8	1.5	3.1	3.5	1.4	1.4	1.8	6.3
Pleasantness	3.9	1.3	3.5	3.5	3.5	1.5	2.4	7.3

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	7.5	8.4	9.0	8.9	7.0	6.5	7.4	10.0
Distinctness	7.7	7.6	7.9	8.6	6.3	5.9	7.4	10.0
Naturalness	3.2	4.1	4.8	5.9	5.9	3.3	5.0	10.0
Pleasantness	3.2	3.2	5.5	6.1	5.8	2.8	2.6	8.3

Subject 2

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	2.6	3.9	9.3	0.8	1.7	2.0	1.6	3.5
Distinctness	3.5	2.3	9.4	1.9	1.9	2.6	0.4	1.8
Naturalness	3.5	2.1	7.9	1.5	1.0	4.8	1.2	3.0
Pleasantness	3.8	3.3	7.9	2.6	2.8	5.2	2.0	2.5

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	5.5	1.2	1.3	4.1	2.2	3.0	2.2	9.1
Distinctness	5.5	0.3	2.1	4.3	1.8	2.5	0.5	9.2
Naturalness	7.2	0.8	2.5	6.8	1.9	2.0	0.7	7.4
Pleasantness	6.9	0.5	2.5	6.8	1.5	2.1	0.8	8.4

Subject 3

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	2.5	2.8	2.8	2.9	3.0	2.8	2.6	8.5
Distinctness	3.3	1.3	2.3	7.5	2.2	4.6	6.8	8.3
Naturalness	3.2	3.2	2.3	6.5	7.2	1.9	7.9	8.3
Pleasantness	5.5	3.5	5.1	5.2	4.2	6.9	3.3	7.5

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	2.9	3.9	3.5	2.3	4.0	4.4	5.0	4.0
Distinctness	4.2	7.0	3.2	2.4	4.7	2.5	3.4	4.7
Naturalness	5.2	7.9	1.4	2.0	5.7	2.8	4.2	5.7
Pleasantness	5.1	5.6	5.7	5.2	5.1	5.0	5.0	5.1

Subject 4

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	8.6	5.2	6.7	7.7	2.1	6.5	2.8	10.0
Distinctness	7.5	3.4	5.1	6.0	2.1	4.6	1.8	10.0
Naturalness	7.6	4.3	5.4	6.0	2.4	5.5	3.3	0.10
Pleasantness	6.8	4.8	3.5	5.0	1.8	5.1	2.6	0.00

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	9.5	5.7	2.3	4.2	0.0	2.2	5.9	8.2
Distinctness	8.6	4.9	1.0	1.2	0.0	2.2	3.6	7.5
Naturalness	8.4	4.6	0.9	0.6	0.1	3.0	4.6	9.4
Pleasantness	8.1	3.5	1.6	0.6	7.4	2.4	4.5	6.9

Subject 5

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	2.7	2.2	3.0	0.7	1.9	4.9	3.0	10.0
Distinctness	1.0	2.0	3.4	3.9	1.8	3.7	0.4	9.8
Naturalness	0.8	0.7	0.1	0.3	0.1	2.1	0.1	10.0
Pleasantness	0.7	0.1	1.5	0.3	0.7	1.0	0.1	10.0

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	1.7	1.3	0.0	5.2	0.1	0.2	3.5	10.0
Distinctness	o.5	0.3	0.0	4.1	0.3	1.2	2.1	9.7
Naturalness	0.2	0.0	0.0	1.2	0.0	0.2	0.2	10.0
Pleasantness	0.5	0.3	0.0	0.1	0.1	0.2	0.9	10.0

Subject 6

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	4.8	5.1	4.6	5.9	7.2	5.4	6.7	9.0
Distinctness	3.4	1.8	3.7	5.3	4.4	4.4	2.7	9.2
Naturalness	2.4	1.5	1.1	4.9	2.6	2.9	2.1	10.0
Pleasantness	2.4	1.6	2.1	4.1	2.7	3.1	1.9	9.7

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	8.5	7.7	6.9	8.7	8.7	8.2	7.2	9.7
Distinctness	8.0	5.6	4.6	6.4	8.2	5.1	3.8	9.5
Naturalness	3.4	3.1	2.9	4.6	2.8	3.8	1.7	10.0
Pleasantness	4.2	2.9	2.8	4.6	2.3	4.6	0.8	10.0

Subject 7

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	5.9	6.5	3.7	5.8	5.2	7.0	1.8	10.0
Distinctness	2.9	4.5	1.1	2.1	3.4	6.5	1.8	10.0
Naturalness	3.0	2.6	1.6	4.9	4.9	5.9	6.7	10.0
Pleasantness	3.4	2.5	1.9	5.2	5.2	6.3	6.7	10.0

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	4.4	6.3	4.9	7.8	7.0	7.2	5.9	10.0
Distinctness	2.1	5.7	3.4	5.7	5.9	5.2	3.5	10.0
Naturalness	3.9	5.0	4.2	2.4	4.1	6.6	3.9	10.0
Pleasantness	4.6	5.2	3.7	3.2	4.6	6.1	4.9	10.0

Subject 8

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	4.3	4.3	2.0	4.6	7.2	4.7	8.9	8.6
Distinctness	3.6	4.1	3.4	4.0	5.0	4.1	7.7	8.8
Naturalness	2.9	3.6	1.8	3.6	3.7	5.6	6.2	10.0
Pleasantness	3.7	4.4	3.3	5.6	5.0	4.5	4.5	7.4

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	7.5	4.6	3.8	3.4	6.3	6.3	7.5	9.7
Distinctness	6.4	1.3	2.7	2.9	4.0	3.4	6.8	9.6
Naturalness	2.6	1.4	0.9	1.8	2.2	2.4	3.5	10.0
Pleasantness	3.5	4.5	3.5	4.5	3.6	5.0	5.0	9.2

Subject 9

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	5.4	2.6	3.7	2.5	2.6	6.5	4.7	10.0
Distinctness	4.7	1.3	4.8	3.8	2.6	5.2	1.9	10.0
Naturalness	4.6	2.5	3.8	3.3	3.5	3.0	3.0	10.0
Pleasantness	3.9	2.3	5.3	2.4	3.6	4.0	3.9	10.0

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	3.7	2.8	1.8	9.0	3.6	6.1	3.4	10.0
Distinctness	3.0	1.1	0.5	8.3	2.6	5.4	3.2	10.0
Naturalness	3.8	2.1	1.6	7.8	3.6	1.4	1.1	10.0
Pleasantness	4.4	2.1	1.6	6.8	4.4	3.1	1.8	10.0

Subject 10

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	4.7	2.6	4.7	4.6	5.1	2.8	2.7	9.6
Distinctness	3.5	1.9	3.0	3.8	2.5	3.2	2.8	9.0
Naturalness	3.8	2.2	0.7	3.4	2.1	3.4	2.2	8.5
Pleasantness	1.5	1.2	0.8	1.3	2.2	4.0	1.5	7.5

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	6.7	5.5	3.2	5.8	7.0	5.4	6.5	10.0
Distinctness	6.6	2.8	2.2	5.0	5.0	3.8	2.1	10.0
Naturalness	2.9	3.2	1.2	3.0	3.2	2.9	3.4	8.8
Pleasantness	2.5	1.9	2.6	4.6	4.2	23	1.5	8.0

Subject 11

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	2.5	2.0	1.8	2.9	4.5	4.5	3.5	8.7
Distinctness	2.6	2.5	1.6	3.5	3.4	3.9	3.0	9.0
Naturalness	3.9	2.1	2.3	2.9	1.9	3.0	4.0	9.1
Pleasantness	4.2	2.0	2.4	2.8	1.9	2.8	4.9	9.2

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	5.5	1.8	3.6	3.9	7.0	3.4	2.4	10.0
Distinctness	5.4	1.8	2.9	3.8	3.4	3.4	3.3	10.0
Naturalness	3.2	2.0	5.7	6.9	2.1	2.1	6.6	10.0
Pleasantness	3.1	2.1	5.7	7.0	2.1	2.4	7.2	10.0

Subject 12

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	3.2	3.6	3.3	8.1	8.8	4.7	5.6	10.0
Distinctness	2.3	2.8	1.8	6.9	3.0	0.2	4.0	10.0
Naturalness	0.2	0.4	0.5	5.2	2.8	1.1	1.6	10.0
Pleasantness	1.3	1.5	2.1	6.8	4.9	1.9	1.7	10.0

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	0.3	2.0	0.9	7.6	9.2	4.6	4.8	10.0
Distinctness	7.5	3.3	0.4	8.4	9.0	3.8	3.5	10.0
Naturalness	5.3	2.1	1.4	3.9	7.0	3.0	2.2	10.0
Pleasantness	4.4	3.4	1.7	5.0	7.6	4.6	2.6	10.0

Subject 13

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	7.8	3.8	7.9	9.3	7.1	5.0	4.7	9.9
Distinctness	8.6	7.6	7.3	8.6	6.8	7.7	2.9	8.8
Naturalness	2.5	1.3	4.1	3.9	2.5	2.9	4.0	9.9
Pleasantness	4.4	2.0	4.5	6.5	2.3	3.2	3.3	8.7

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	8.0	4.0	4.8	6.8	6.7	7.1	4.2	9.3
Distinctness	7.4	5.8	5.9	7.1	3.4	6.4	3.5	9.7
Naturalness	4.4	1.9	1.2	4.7	3.8	4.4	3.7	10.0
Pleasantness	3.8	3.0	3.3	6.0	2.9	3.7	5.6	9.6

Subject 14

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	3.7	2.3	5.0	7.0	3.2	4.3	53	9.7
Distinctness	1.1	0.5	2.5	6.9	1.9	2.6	2.3	9.6
Naturalness	3.4	2.0	1.2	5.0	2.0	2.8	5.5	9.5
Pleasantness	1.8	1.2	1.5	5.0	0.9	2.0	6.7	9.6

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	5.2	2.0	3.3	3.2	5.5	1.3	4.0	9.8
Distinctness	2.4	2.0	1.9	2.2	4.0	1.4	5.8	9.9
Naturalness	5.0	3.7	2.9	5.4	4.2	3.7	5.8	9.8
Pleasantness	3.8	1.0	1.2	4.0	4.3	28	5.8	8.8

Subject 15

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	5.1	4.2	3.1	5.3	3.8	4.4	6.2	9.7
Distinctness	2.1	3.8	1.1	4.5	2.3	2.4	4.6	9.7
Naturalness	1.8	3.4	2.3	4.6	1.0	2.3	4.6	10.0
Pleasantness	3.9	4.1	2.2	5.8	5.5	2.3	7.2	7.7

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	7.0	7.0	4.6	6.5	6.7	4.4	5.2	9.8
Distinctness	5.4	7.0	2.4	3.9	4.7	4.6	3.1	9.8
Naturalness	4.4	3.8	2.6	3.9	5.3	3.1	5.4	9.8
Pleasantness	3.7	3.9	2.7	5.4	5.5	4.9	3.6	9.8

Subject 16

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	2.9	5.2	1.8	3.2	3.7	7.7	1.7	10.0
Distinctness	2.8	3.9	1.4	2.7	2.4	6.0	1.7	10.0
Naturalness	1.8	2.9	1.4	4.7	3.4	1.1	3.9	10.0
Pleasantness	1.8	5.5	1.6	2.9	4.9	1.1	2.4	10.0

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	5.7	5.4	1.8	6.6	2.6	4.9	5.7	10.0
Distinctness	4.7	0.6	1.1	4.2	2.6	3.4	5.7	10.0
Naturalness	3.1	0.4	0.6	7.1	1.5	3.1	0.8	10.0
Pleasantness	5.2	3.4	1.5	6.5	3.3	4.9	0.8	10.0

Subject 17

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	4.0	1.0	1.5	3.0	1.9	2.0	2.7	9.3
Distinctness	3.5	3.1	1.1	3.6	4.3	1.7	4.9	9.0
Naturalness	3.0	2.2	1.0	3.8	1.5	1.3	1.7	10.0
Pleasantness	3.7	1.9	0.7	5.9	4.0	2.4	5.6	10.0

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	6.3	3.0	2.0	4.9	1.7	1.1	2.7	10.0
Distinctness	5.6	1.7	2.9	5.2	1.2	2.1	3.1	10.0
Naturalness	4.7	1.5	1.1	3.5	1.1	1.3	2.3	10.0
Pleasantness	5.3	2.3	2.3	5.2	1.8	2.1	3.9	10.0

Subject 18

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	8.3	9.1	8.3	9.3	5.0	6.8	8.3	10.0
Distinctness	3.5	5.5	4.3	8.4	2.2	4.8	3.3	10.0
Naturalness	0.7	2.5	0.7	1.3	0.7	1.2	1.2	10.0
Pleasantness	2.7	1.2	2.0	5.3	3.0	1.3	29	10.0

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	8.8	8.8	7.8	9.8	7.6	7.3	8.8	10.0
Distinctness	4.3	4.5	2.3	7.0	1.2	5.0	3.2	10.0
Naturalness	1.2	1.0	0.7	2.0	0.4	0.8	0.9	10.0
Pleasantness	2.8	2.5	1.9	3.0	3.2	3.5	2.2	10.0

Subject 19

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	6.0	5.3	5.9	5.4	7.2	6.5	5.6	9.7
Distinctness	3.3	6.2	3.3	6.0	6.6	2.6	6.4	9.8
Naturalness	3.3	4.8	2.8	4.8	3.4	3.3	2.8	7.4
Pleasantness	2.9	3.9	2.2	3.9	3.4	2.1	2.7	8.4

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	7.0	6.0	8.2	8.1	6.8	6.7	8.0	9.5
Distinctness	3.3	3.1	2.5	3.7	4.6	2.6	7.3	9.4
Naturalness	3.3	2.2	3.3	3.4	3.0	3.5	4.2	6.5
Pleasantness	2.9	2.5	2.7	2.7	3.2	3.3	3.9	7.2

Subject 20

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	6.5	5.2	5.5	6.0	3.5	5.2	5.9	9.5
Distinctness	6.3	6.3	6.3	7.6	4.6	5.7	6.2	10.0
Naturalness	2.1	2.2	1.8	2.4	1.8	2.8	2.5	10.0
Pleasantness	3.0	2.8	3.3	2.6	2.1	2.6	2.3	10.0

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	2.0	6.4	7.9	6.3	3.8	6.4	6.8	10.0
Distinctness	2.4	6.3	8.0	5.7	3.5	6.8	3.9	10.0
Naturalness	2.8	2.0	2.1	1.8	1.8	1.4	1.3	10.0
Pleasantness	2.5	2.1	2.3	2.3	1.8	1.8	2.2	10.0

Subject 21

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	6.2	6.2	4.8	3.9	6.8	7.3	5.4	4.8
Distinctness	6.2	6.3	4.4	3.7	5.7	7.4	4.7	4.4
Naturalness	5.1	4.3	5.2	2.9	5.0	4.5	3.5	5.2
Pleasantness	4.3	4.2	2.4	1.4	2.8	4.8	1.5	2.4

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	7.2	8.6	7.0	9.0	6.8	8.0	7.7	10.0
Distinctness	7.4	7.2	3.2	5.4	7.8	6.8	9.0	10.0
Naturalness	7.0	4.0	3.9	8.0	6.2	4.4	6.2	9.5
Pleasantness	3.5	3.0	1.8	3.2	6.2	3.3	2.5	10.0

Subject 22

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	6.7	2.6	5.2	7.3	3.0	5.2	4.0	9.9
Distinctness	5.7	2.5	3.7	6.5	4.0	5.0	4.7	9.6
Naturalness	2.2	1.2	2.2	8.3	3.2	3.3	7.1	8.8
Pleasantness	2.6	2.0	2.7	7.7	5.2	3.9	2.5	8.6

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	2.7	1.7	6.0	7.0	6.0	3.8	4.8	9.3
Distinctness	2.7	1.9	5.0	8.1	5.3	2.7	5.2	9.2
Naturalness	5.8	3.2	5.0	4.0	4.0	3.5	3.8	9.6
Pleasantness	3.5	2.0	2.7	5.5	1.2	1.4	3.7	9.6

Subject 23

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	8.4	7.3	5.6	4.4	8.7	5.8	5.5	10.0
Distinctness	2.0	2.6	3.0	3.8	3.2	3.2	3.9	9.6
Naturalness	2.1	1.5	1.7	1.9	1.6	2.8	2.2	8.5
Pleasantness	1.8	1.3	2.1	1.4	2.0	2.2	2.0	8.8

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	3.9	3.3	3.0	5.3	3.7	3.4	4.6	9.9
Distinctness	1.9	3.3	1.7	5.1	3.5	1.7	4.3	9.8
Naturalness	1.8	2.4	1.4	3.0	1.7	1.8	1.5	8.3
Pleasantness	1.7	2.1	1.1	2.0	1.8	1.2	1.8	9.4

Subject 24

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	4.4	1.8	3.1	3.8	3.5	2.4	2.7	9.5
Distinctness	3.5	2.5	2.1	3.7	3.4	2.5	3.4	9.0
Naturalness	3.4	2.2	0.9	3.6	3.3	4.4	2.0	9.5
Pleasantness	2.6	1.6	0.8	3.6	3.1	3.2	3.6	8.8

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7	S8
Intelligibility	6.5	4.3	2.6	5.4	4.4	3.3	4.6	10.0
Distinctness	6.1	2.3	2.6	5.1	3.1	3.0	2.6	10.0
Naturalness	3.8	2.4	1.2	3.3	2.2	2.1	2.9	9.5
Pleasantness	3.9	2.1	2.5	4.9	3.0	2.2	2.7	9.0

A.4.2.2 First MDS-SDS Test Multidimensional Scaling Data

Subject 1

	S2	S3	S4	S5	S6	S7	S8
S1	3.90	4.59	6.30	5.47	0.10	0.20	6.60
S2		5.65	4.20	5.15	2.00	1.67	7.10
S3			7.12	0.70	1.25	6.50	4.74
S4				5.20	4.00	3.10	3.36
S5					5.00	5.72	9.30
S6						10.00	5.40
S7							5.64

Subject 2

	S2	S3	S4	S5	S6	S7	S8
S1	3.10	3.03	3.50	4.28	2.80	2.82	4.60
S2		4.61	2.90	2.64	1.20	0.61	8.90
S3			9.42	0.60	1.25	3.00	2.77
S4				3.50	2.91	1.70	1.90
S5					2.10	2.50	9.70
S6						10.00	3.50
S7							3.55

Subject 5

	S2	S3	S4	S5	S6	S7	S8
S1	6.60	5.03	3.30	4.98	0.10	0.68	6.10
S2		4.23	2.50	3.98	2.00	2.38	10.00
S3			10.00	0.00	0.19	4.80	4.46
S4				2.20	1.60	1.10	1.05
S5					5.10	7.31	10.00
S6						10.00	4.80
S7							4.41

Subject 6

	S2	S3	S4	S5	S6	S7	S8
S1	1.80	1.78	4.20	3.01	1.10	1.69	1.60
S2		3.01	4.10	3.50	1.30	1.50	9.50
S3			8.83	0.60	1.23	1.40	1.38
S4				2.30	1.65	1.50	1.94
S5					2.60	3.04	9.20
S6						9.64	3.30
S7							2.15

Subject 7

	S2	S3	S4	S5	S6	S7	S8
S1	4.80	4.45	7.20	5.61	0.00	0.65	2.00
S2		3.52	3.30	4.45	2.30	2.50	9.50
S3			9.16	0.00	1.41	5.90	4.97
S4				0.50	1.91	3.20	10.00
S5					1.90	3.98	10.00
S6						9.83	4.90
S7							6.18

Subject 8

	S2	S3	S4	S5	S6	S7	S8
S1	6.10	5.31	7.80	6.64	0.10	0.95	4.00
S2		4.26	3.20	4.07	1.50	1.51	9.00
S3			9.05	0.70	1.35	3.50	4.36
S4				1.20	1.05	2.60	1.36
S5					4.80	6.73	9.40
S6						9.18	4.70
S7							5.69

Subject 9

	S2	S3	S4	S5	S6	S7	S8
S1	7.40	5.66	6.70	7.05	1.90	1.56	5.60
S2		4.74	2.60	4.88	0.50	0.89	9.50
S3			9.45	0.70	1.40	7.10	6.22
S4				1.30	1.31	0.80	0.88
S5					2.70	6.03	9.80
S6						9.45	7.40
S7							7.54

Subject 10

	S2	S3	S4	S5	S6	S7	S8
S1	8.70	8.55	8.20	8.36	0.60	1.45	6.40
S2		7.28	8.60	8.07	2.50	2.76	9.70
S3			8.93	0.30	0.67	9.00	9.16
S4				5.70	5.10	5.98	4.81
S5					8.60	7.84	9.70
S6						10.00	8.50
S7							9.07

Subject 11

	S2	S3	S4	S5	S6	S7	S8
S1	3.90	4.59	6.30	5.47	0.10	0.20	6.60
S2		5.65	4.20	5.15	2.00	1.67	7.10
S3			7.12	0.70	1.25	6.50	4.74
S4				5.20	4.00	3.10	3.36
S5					5.00	5.72	9.30
S6						10.00	5.40
S7							5.64

Subject 12

	S2	S3	S4	S5	S6	S7	S8
S1	6.90	7.72	7.70	7.39	1.00	0.00	3.70
S2		3.81	3.40	2.75	2.80	2.81	9.70
S3			9.96	0.10	0.65	7.90	7.50
S4				2.70	4.38	7.70	4.97
S5					9.60	10.00	9.90
S6						9.92	6.60
S7							7.11

Subject 14

	S2	S3	S4	S5	S6	S7	S8
S1	8.30	7.10	2.30	5.61	0.60	1.67	7.80
S2		5.33	7.40	6.80	3.00	2.04	9.00
S3			78.19	0.20	2.02	7.40	8.43
S4				8.20	6.37	8.50	6.48
S5					7.10	8.93	9.90
S6						7.64	8.30
S7							6.74

Subject 15

	S2	S3	S4	S5	S6	S7	S8
S1	2.90	3.20	5.30	4.04	0.30	0.20	1.80
S2		1.74	2.60	3.52	2.10	1.82	9.20
S3			8.87	0.30	1.07	3.10	3.32
S4				3.10	1.80	2.10	1.38
S5					2.60	3.83	9.90
S6						10.00	2.60
S7							4.13

Subject 16

	S2	S3	S4	S5	S6	S7	S8
S1	5.70	4.73	3.90	5.25	0.60	0.78	4.90
S2		4.90	3..70	3.94	0.90	1.35	9.50
S3			9.44	0.10	0.52	5.40	5.49
S4				2.80	2.31	2.60	2.36
S5					4.40	5.89	9.20
S6						8.34	6.90
S7							6.01

Subject 17

	S2	S3	S4	S5	S6	S7	S8
S1	9.00	8.35	8.60	8.64	3.80	3.30	7.30
S2		5.45	5.00	4.76	3.00	2.71	9.50
S3			10.00	1.80	1.05	7.80	6.19
S4				8.50	9.78	7.50	8.82
S5					7.60	7.56	10.00
S6						8.76	7.50
S7							6.82

Subject 20

	S2	S3	S4	S5	S6	S7	S8
S1	8.80	7.72	8.00	8.56	0.00	1.06	7.50
S2		7.35	7.50	7.20	1.30	0.71	9.90
S3			9.69	0.00	1.04	5.80	6.72
S4				2.30	1.21	0.60	0.69
S5					6.70	7.84	9.90
S6						10.00	9.50
S7							7.59

Subject 21

	S2	S3	S4	S5	S6	S7	S8
S1	3.70	5.06	6.00	5.58	0.00	1.53	4.40
S2		4.09	3.70	4.31	4.20	3.30	9.40
S3			10.00	0.00	1.15	7.80	6.52
S4				1.30	1.22	3.80	2.14
S5					4.20	5.74	9.70
S6						8.54	7.30
S7							7.09

Subject 22

	S2	S3	S4	S5	S6	S7	S8
S1	5.80	5.19	6.60	5.80	2.20	1.37	7.40
S2		7.28	5.20	6.69	2.00	2.54	7.10
S3			6.44	1.00	0.72	6.30	4.92
S4				2.20	3.17	3.20	1.99
S5					3.00	4.39	8.40
S6						8.57	5.80
S7							5.43

Subject 23

	S2	S3	S4	S5	S6	S7	S8
S1	4.40	5.14	4.80	5.77	0.60	1.46	5.90
S2		4.31	3.60	3.42	1.10	2.15	5.40
S3			6.51	0.00	1.23	6.10	4.88
S4				2.30	2.14	2.70	2.73
S5					4.00	5.96	9.90
S6						8.71	5.70
S7							5.77

Subject 24

	S2	S3	S4	S5	S6	S7	S8
S1	8.90	7.60	8.40	7.63	2.20	1.19	6.90
S2		5.90	5.80	5.47	2.80	2.66	9.60
S3			10.00	1.10	0.69	8.40	8.78
S4				7.20	8.39	6.70	7.64
S5					8.20	7.18	9.90
S6						7.65	8.00
S7							8.81

A.4.3 Second MDS-SDS Test Multidimensional Scaling and Semantic Differential Scaling Data

A.4.3.1 Second MDS-SDS Test Semantic Differential Scaling- Data

Subject 1

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7
Intelligibility	4.9	2.1	4.3	2.7	2.0	4.8	2.9
Distinctness	4.1	3.3	4.3	2.8	1.4	6.5	3.4
Naturalness	3.8	1.5	3.1	3.5	1.4	1.4	1.8
Pleasantness	3.9	1.3	3.5	3.5	3.5	1.5	2.4
Listening effort	3.2	4.2	2.1	3.2	2.2	3.4	7.5

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7
Intelligibility	7.5	8.4	9.0	8.9	7.0	6.5	10.0
Distinctness	7.7	7.6	7.9	8.6	6.3	5.9	10.0
Naturalness	3.2	4.1	4.8	5.9	5.9	3.3	10.0
Pleasantness	3.2	3.2	5.5	6.1	5.8	2.8	8.3
Listening effort	3.2	4.2	2.1	3.2	2.2	3.4	9.5

Subject 2

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7
Intelligibility	2.6	3.9	9.3	0.8	1.7	2.0	3.5
Distinctness	3.5	2.3	9.4	1.9	1.9	2.6	1.8
Naturalness	3.5	2.1	7.9	1.5	1.0	4.8	3.0
Pleasantness	3.8	3.3	7.9	2.6	2.8	5.2	2.5
Listening effort	3.2	4.2	2.1	3.2	2.2	3.4	3.5

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7
Intelligibility	5.5	1.2	1.3	4.1	2.2	3.0	9.1
Distinctness	5.5	0.3	2.1	4.3	1.8	2.5	9.2
Naturalness	7.2	0.8	2.5	6.8	1.9	2.0	7.4
Pleasantness	6.9	0.5	2.5	6.8	1.5	2.1	8.4
Listening effort	3.2	4.2	2.1	3.2	2.2	3.4	7.5

Subject 3

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7
Intelligibility	2.5	2.8	2.8	2.9	3.0	2.8	8.5
Distinctness	3.3	1.3	2.3	7.5	2.2	4.6	8.3
Naturalness	3.2	3.2	2.3	6.5	7.2	1.9	8.3
Pleasantness	5.5	3.5	5.1	5.2	4.2	6.9	7.5
Listening effort	3.2	4.2	2.1	3.2	2.2	3.4	7.5

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7
Intelligibility	2.9	3.9	3.5	2.3	4.0	4.4	4.0
Distinctness	4.2	7.0	3.2	2.4	4.7	2.5	4.7
Naturalness	5.2	7.9	1.4	2.0	5.7	2.8	4.2
Pleasantness	5.1	5.6	5.7	5.2	5.1	5.0	5.0
Listening effort	3.2	4.2	2.1	3.2	2.2	3.4	7.5

Subject 4

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7
Intelligibility	8.6	5.2	6.7	7.7	2.1	6.5	10.0
Distinctness	7.5	3.4	5.1	6.0	2.1	4.6	10.0
Naturalness	7.6	4.3	5.4	6.0	2.4	5.5	0.10
Pleasantness	6.8	4.8	3.5	5.0	1.8	5.1	0.00
Listening effort	3.2	4.2	2.1	3.2	2.2	3.4	7.5

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7
Intelligibility	9.5	5.7	2.3	4.2	0.0	2.2	8.2
Distinctness	8.6	4.9	1.0	1.2	0.0	2.2	7.5
Naturalness	8.4	4.6	0.9	0.6	0.1	3.0	9.4
Pleasantness	8.1	3.5	1.6	0.6	7.4	2.4	6.9
Listening effort	3.2	4.2	2.1	3.2	2.2	3.4	7.5

Subject 5

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7
Intelligibility	2.7	2.2	3.0	0.7	1.9	4.9	10.0
Distinctness	1.0	2.0	3.4	3.9	1.8	3.7	9.8
Naturalness	0.8	0.7	0.1	0.3	0.1	2.1	10.0
Pleasantness	0.7	0.1	1.5	0.3	0.7	1.0	10.0
Listening effort	3.2	4.2	2.1	3.2	2.2	3.4	10.0

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7
Intelligibility	1.7	1.3	0.0	5.2	0.1	0.2	10.0
Distinctness	o.5	0.3	0.0	4.1	0.3	1.2	9.7
Naturalness	0.2	0.0	0.0	1.2	0.0	0.2	10.0
Pleasantness	0.5	0.3	0.0	0.1	0.1	0.2	10.0
Listening effort	3.2	4.2	2.1	3.2	2.2	3.4	10.0

Subject 6

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7
Intelligibility	4.8	5.1	4.6	5.9	7.2	5.4	9.0
Distinctness	3.4	1.8	3.7	5.3	4.4	4.4	9.2
Naturalness	2.4	1.5	1.1	4.9	2.6	2.9	10.0
Pleasantness	2.4	1.6	2.1	4.1	2.7	3.1	9.7
Listening effort	3.2	4.2	2.1	3.2	2.2	3.4	9.5

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7
Intelligibility	8.5	7.7	6.9	8.7	8.7	8.2	9.7
Distinctness	8.0	5.6	4.6	6.4	8.2	5.1	9.5
Naturalness	3.4	3.1	2.9	4.6	2.8	3.8	10.0
Pleasantness	4.2	2.9	2.8	4.6	2.3	4.6	10.0
Listening effort	3.2	4.2	2.1	3.2	2.2	3.4	10.0

Subject 7

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7
Intelligibility	5.9	6.5	3.7	5.8	5.2	7.0	10.0
Distinctness	2.9	4.5	1.1	2.1	3.4	6.5	10.0
Naturalness	3.0	2.6	1.6	4.9	4.9	5.9	10.0
Pleasantness	3.4	2.5	1.9	5.2	5.2	6.3	10.0
Listening effort	3.2	4.2	2.1	3.2	2.2	3.4	10.0

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7
Intelligibility	4.4	6.3	4.9	7.8	7.0	7.2	10.0
Distinctness	2.1	5.7	3.4	5.7	5.9	5.2	10.0
Naturalness	3.9	5.0	4.2	2.4	4.1	6.6	10.0
Pleasantness	4.6	5.2	3.7	3.2	4.6	6.1	10.0
Listening effort	3.2	4.2	2.1	3.2	2.2	3.4	10.0

Subject 8

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7
Intelligibility	4.3	4.3	2.0	4.6	7.2	4.7	8.9
Distinctness	3.6	4.1	3.4	4.0	5.0	4.1	7.7
Naturalness	2.9	3.6	1.8	3.6	3.7	5.6	6.2
Pleasantness	3.7	4.4	3.3	5.6	5.0	4.5	4.5
Listening effort	3.2	4.2	2.1	3.2	2.2	3.4	7.5

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7
Intelligibility	7.5	4.6	3.8	3.4	6.3	6.3	9.7
Distinctness	6.4	1.3	2.7	2.9	4.0	3.4	9.6
Naturalness	2.6	1.4	0.9	1.8	2.2	2.4	10.0
Pleasantness	3.5	4.5	3.5	4.5	3.6	5.0	9.2
Listening effort	3.2	4.2	2.1	3.2	2.2	3.4	7.5

Subject 9

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7
Intelligibility	5.4	2.6	3.7	2.5	2.6	6.5	10.0
Distinctness	4.7	1.3	4.8	3.8	2.6	5.2	10.0
Naturalness	4.6	2.5	3.8	3.3	3.5	3.0	10.0
Pleasantness	3.9	2.3	5.3	2.4	3.6	4.0	10.0
Listening effort	3.2	4.2	2.1	3.2	2.2	3.4	10.0

Second Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7
Intelligibility	3.7	2.8	1.8	9.0	3.6	6.1	10.0
Distinctness	3.0	1.1	0.5	8.3	2.6	5.4	10.0
Naturalness	3.8	2.1	1.6	7.8	3.6	1.4	10.0
Pleasantness	4.4	2.1	1.6	6.8	4.4	3.1	10.0
Listening effort	3.2	4.2	2.1	3.2	2.2	3.4	10.0

Subject 10

First Semantic Differential Scaling Test

Treatments	S1	S2	S3	S4	S5	S6	S7
Intelligibility	4.7	2.6	4.7	4.6	5.1	2.8	9.6
Distinctness	3.5	1.9	3.0	3.8	2.5	3.2	9.0
Naturalness	3.8	2.2	0.7	3.4	2.1	3.4	8.5
Pleasantness	1.5	1.2	0.8	1.3	2.2	4.0	7.5
Listening effort	3.2	4.2	2.1	3.2	2.2	3.4	7.5

A.4.3.2 Second MDS-SDS Test Multidimensional Scaling Data

Subject 1

	S2	S3	S4	S5	S6	S7
S1	1.70	9.60	9.70	9.90	3.50	9.70
S2		9.70	9.70	9.60	1.90	9.70
S3			7.40	1.70	9.80	7.70
S4				1.70	9.70	5.20
S5					9.60	4.70
S6						9.60

Subject 2

	S2	S3	S4	S5	S6	S7
S1	0.30	9.70	1.00	8.90	0.30	8.70
S2		8.70	9.70	8.40	0.80	8.20
S3			5.70	0.30	9.00	4.50
S4				0.00	9.70	4.20
S5					9.60	4.10
S6						9.40

Subject 3

	S2	S3	S4	S5	S6	S7
S1	0.80	7.70	8.70	3.40	1.50	6.40
S2		9.20	8.00	8.10	2.60	9.20
S3			1.60	1.60	8.40	6.40
S4				0.60	8.20	5.70
S5					9.00	3.90
S6						9.90

Subject 4

	S2	S3	S4	S5	S6	S7
S1	2.50	3.20	4.10	4.00	0.50	2.70
S2		4.20	5.80	6.00	4.20	6.80
S3			1.90	0.40	3.00	0.60
S4				0.00	4.10	1.60
S5					7.00	0.90
S6						5.90

Subject 5		S2	S3	S4	S5	S6	S7
	S1	1.10	5.10	7.90	8.10	1.50	7.20
	S2		8.70	9.00	4.10	1.80	6.60
	S3			2.30	0.80	7.30	3.10
	S4				0.30	8.60	3.00
	S5					8.40	4.30
	S6						8.70

Subject 6		S2	S3	S4	S5	S6	S7
	S1	1.50	5.70	6.50	3.40	1.20	4.80
	S2		6.50	6.70	7.00	3.30	6.80
	S3			1.40	1.20	5.70	3.50
	S4				0.50	6.40	3.60
	S5					7.90	2.50
	S6						7.80

Subject 7		S2	S3	S4	S5	S6	S7
	S1	2.10	8.40	10.00	8.40	1.20	10.00
	S2		10.00	9.20	10.00	0.70	10.00
	S3			3.30	8.80	10.00	4.80
	S4				0.00	10.00	6.50
	S5					10.00	2.90
	S6						10.00

Subject 8		S2	S3	S4	S5	S6	S7
	S1	1.30	6.20	5.40	5.40	1.80	9.70
	S2		5.10	6.28	8.50	4.60	7.20
	S3			2.60	2.20	7.50	4.60
	S4				0.20	6.40	4.00
	S5					6.90	4.40
	S6						9.80

Subject 9		S2	S3	S4	S5	S6	S7
	S1	2.10	6.80	5.70	5.20	2.00	6.50
	S2		6.30	6.80	7.90	3.10	7.40
	S3			3.40	2.50	7.70	3.90
	S4				0.40	6.90	3.80
	S5					6.50	4.00
	S6						9.00

Subject 10		S2	S3	S4	S5	S6	S7
	S1	1.70	5.50	6.40	3.70	1.00	4.60
	S2		6.70	6.80	7.10	3.40	7.00
	S3			1.30	1.00	5.70	3.50
	S4				0.30	6.20	3.70
	S5					8.00	2.40
	S6						7.90

A.4.4 TFW Parameter Listening Effort Test

A.4.4.1 TFW Parameter Listening Effort Test Rep. 1

Subject 1											
Level 1	4	1	2	1	3	1	1	2	3	1	0
Level 2	3	0	1	2	3	2	1	2	3	1	0
Level 3	4	0	1	2	3	2	1	2	3	1	0
Level 4	4	2	1	2	3	2	1	1	3	0	0
Level 5	4	0	2	2	3	2	1	1	3	1	0
Subject 2											
Level 1	4	2	1	2	4	2	2	1	3	2	0
Level 2	4	3	2	1	3	4	0	2	3	1	1
Level 3	4	4	2	3	4	3	1	1	2	1	0
Level 4	4	1	2	3	4	3	1	3	2	1	0
Level 5	4	0	1	2	4	2	0	2	4	1	0
Subject 3											
Level 1	4	1	2	2	4	4	3	3	4	2	0
Level 2	4	1	1	2	4	4	3	3	2	3	0
Level 3	4	1	0	3	4	4	2	3	3	1	0
Level 4	4	0	0	1	4	3	4	2	4	1	0
Level 5	4	0	0	3	4	4	3	2	4	1	0
Subject 4											
Level 1	4	0	1	1	2	4	2	2	2	1	0
Level 2	3	1	1	2	2	3	1	1	2	1	0
Level 3	4	1	0	1	3	3	1	2	2	1	0
Level 4	4	1	2	1	3	3	1	2	2	0	0
Level 5	3	1	0	2	2	2	1	1	3	3	1
Subject 5											
Level 1	3	1	0	2	3	3	2	0	2	0	0
Level 2	3	2	1	0	4	3	0	1	3	1	0
Level 3	4	2	0	1	2	3	1	2	3	1	0
Level 4	4	0	2	2	3	3	2	1	4	0	0
Level 5	4	1	0	1	4	2	2	1	2	0	0
Subject 6											
Level 1	3	2	2	2	3	1	1	0	2	0	0
Level 2	4	2	1	2	3	3	1	2	2	0	0
Level 3	4	1	1	3	3	2	1	1	3	0	0
Level 4	4	2	3	2	3	3	1	2	2	0	1
Level 5	3	1	1	2	3	3	1	1	2	0	0
Subject 7											
Level 1	4	0	2	2	3	3	1	2	3	1	0
Level 2	4	1	2	2	3	3	2	3	3	0	0
Level 3	4	2	2	2	3	3	1	3	3	1	0
Level 4	4	1	1	2	3	3	1	2	4	0	0
Level 5	4	1	1	2	3	3	1	2	2	1	0

Subject 8											
Level 1	3	1	0	0	3	3	1	2	2	1	0
Level 2	4	1	0	1	2	1	2	2	3	2	0
Level 3	4	1	2	2	2	1	2	1	3	1	0
Level 4	4	2	1	2	3	3	2	2	2	1	0
Level 5	4	0	2	1	4	1	1	2	3	1	1

Subject 9											
Level 1	2	1	1	1	0	0	0	2	2	0	1
Level 2	3	0	0	3	1	0	0	1	2	1	0
Level 3	3	1	0	2	2	0	1	2	1	1	1
Level 4	4	2	0	0	1	1	1	3	3	0	3
Level 5	4	3	1	1	2	1	1	3	2	1	3

Subject 10											
Level 1	4	1	1	1	2	1	1	2	3	0	0
Level 2	4	0	1	2	3	1	1	2	3	1	0
Level 3	4	0	1	1	3	1	1	1	3	1	0
Level 4	4	9	1	2	3	2	1	3	3	1	0
Level 5	4	1	2	0	3	1	1	3	2	0	0

Subject 11											
Level 1	3	0	0	1	3	2	1	0	2	1	0
Level 2	3	1	0	2	3	1	2	0	3	1	0
Level 3	3	1	0	1	3	1	2	0	3	0	0
Level 4	4	1	0	1	3	1	2	1	2	1	0
Level 5	3	1	0	2	3	1	2	0	3	1	0

A.4.4.2 TFW Parameter Listening Effort Test Rep. 2

Subject 1											
Level 1	4	1	2	2	4	2	2	2	3	1	0
Level 2	4	0	1	3	4	3	1	2	3	1	0
Level 3	4	0	0	2	4	2	2	1	4	1	1
Level 4	4	1	1	2	4	3	1	1	4	1	0
Level 5	4	1	1	2	4	2	1	1	4	2	0

Subject 2											
Level 1	4	1	2	1	4	4	0	0	2	0	0
Level 2	4	3	3	1	4	4	0	1	4	1	1
Level 3	4	2	1	0	4	3	1	3	1	0	0
Level 4	4	1	2	1	3	4	1	1	0	1	0
Level 5	4	0	0	1	4	3	0	2	2	0	0

Subject 3											
Level 1	4	2	2	2	3	2	1	2	4	1	0
Level 2	4	1	3	3	4	3	2	2	3	1	0
Level 3	4	1	1	3	4	3	2	3	3	1	0
Level 4	4	1	1	2	4	3	2	2	2	2	0
Level 5	4	1	2	3	3	2	2	1	4	0	0

Subject 4											
Level 1	4	1	1	1	3	3	2	3	3	1	0
Level 2	4	1	2	2	3	3	1	3	2	1	0
Level 3	4	1	1	1	3	4	1	2	3	0	1
Level 4	4	2	2	1	2	4	0	2	2	0	0
Level 5	4	2	0	3	2	4	0	2	3	2	1

Subject 5

Level 1	3	1	0	2	4	1	1	1	3	1	0
Level 2	4	3	3	1	4	3	0	3	4	1	0
Level 3	3	2	0	1	4	2	0	2	4	1	0
Level 4	4	1	3	3	4	3	2	2	4	2	1
Level 5	3	3	1	1	4	1	1	4	3	2	1

Subject 6

Level 1	4	2	2	2	4	2	1	1	4	1	0
Level 2	4	3	2	3	4	3	1	4	3	1	0
Level 3	4	1	2	2	4	3	2	1	4	2	0
Level 4	4	2	3	3	3	4	2	3	3	1	0
Level 5	4	1	2	3	4	4	1	2	4	0	0

Subject 7

Level 1	4	0	1	3	4	2	1	3	4	1	0
Level 2	4	1	1	4	4	3	4	2	4	1	0
Level 3	4	3	2	4	4	3	3	3	3	0	0
Level 4	4	2	1	3	4	3	1	2	4	0	0
Level 5	4	3	2	3	4	2	0	2	3	3	1

Subject 8

Level 1	4	2	1	0	4	3	1	3	4	2	1
Level 2	4	2	1	1	4	4	3	2	4	2	0
Level 3	4	2	3	2	4	3	2	3	3	1	0
Level 4	3	3	2	3	4	4	2	2	4	1	0
Level 5	4	1	2	2	4	3	1	3	4	2	1

Subject 9

Level 1	4	2	0	2	4	3	0	3	4	0	0
Level 2	4	0	1	3	4	2	1	2	3	0	0
Level 3	4	2	1	1	4	3	3	2	2	1	0
Level 4	4	2	1	0	4	3	3	3	3	0	0
Level 5	4	2	2	0	3	4	3	2	3	1	0

Subject 10

Level 1	4	1	1	0	3	2	2	4	4	1	1
Level 2	4	1	2	1	4	2	3	3	4	2	2
Level 3	4	0	2	0	4	2	3	2	4	1	1
Level 4	4	1	1	1	4	2	2	3	4	2	2
Level 5	4	2	1	0	4	3	2	3	3	1	0

Subject 11

Level 1	3	1	2	2	3	1	3	2	3	1	0
Level 2	3	1	1	2	3	1	2	1	3	2	0
Level 3	3	0	2	3	3	0	0	1	2	1	0
Level 4	3	0	2	2	2	0	0	1	2	1	0
Level 5	3	0	2	2	3	0	0	1	3	2	0

A.4.5 Comparison Listening Effort Tests

A.4.5.1 Comparison Listening Effort Test 1 Rep. 1

Subject 1											
Level 1	4	1	0	1	4	0	0	0	0	1	2
Level 2	4	1	1	3	2	3	2	4	3	2	2
Level 3	4	1	1	1	4	0	0	0	0	1	3
Level 4	4	1	1	1	4	0	0	0	0	0	2
Level 5	3	1	0	1	3	0	0	0	0	1	2
Subject 2											
Level 1	4	3	1	2	4	1	1	2	2	1	3
Level 2	4	3	2	3	4	0	1	2	1	1	4
Level 3	4	3	3	3	4	0	0	0	1	4	3
Level 4	4	2	2	4	4	0	1	2	1	2	3
Level 5	4	3	1	3	4	1	1	2	1	3	4
Subject 3											
Level 1	3	2	3	2	4	1	1	1	1	2	2
Level 2	4	4	2	3	4	1	2	0	1	2	2
Level 3	4	3	3	2	4	0	1	2	2	3	3
Level 4	4	2	2	3	4	0	1	0	1	3	3
Level 5	4	3	4	1	4	0	1	2	1	4	2
Subject 4											
Level 1	4	0	1	1	2	0	0	1	1	2	2
Level 2	4	1	1	1	3	0	3	1	1	1	3
Level 3	4	0	1	1	4	0	0	1	0	1	3
Level 4	4	1	0	1	3	0	0	0	0	0	1
Level 5	3	1	0	1	3	0	0	0	1	1	2
Subject 5											
Level 1	4	1	1	2	4	1	0	1	1	1	4
Level 2	4	1	0	2	4	1	2	0	0	1	3
Level 3	4	2	2	1	4	0	0	0	1	2	4
Level 4	4	0	0	3	4	0	1	0	0	1	4
Level 5	4	1	1	0	3	0	0	0	0	1	4
Subject 6											
Level 1	4	2	2	2	4	0	0	2	2	2	3
Level 2	3	2	2	1	4	1	1	1	1	2	3
Level 3	4	2	1	2	4	0	0	1	2	2	2
Level 4	4	1	1	3	3	0	0	1	1	3	2
Level 5	4	1	1	1	3	1	0	1	1	1	1
Subject 7											
Level 1	4	2	1	1	4	0	0	0	0	1	2
Level 2	4	3	1	2	4	0	0	0	1	1	3
Level 3	3	1	2	2	4	0	0	0	1	1	2
Level 4	3	1	1	1	3	0	0	0	1	1	3
Level 5	2	0	1	0	3	0	0	0	0	1	1

Subject 8											
Level 1	4	1	1	0	3	0	0	0	0	1	1
Level 2	4	1	2	1	4	0	0	0	0	0	2
Level 3	4	0	0	1	4	0	0	0	0	0	2
Level 4	3	0	1	1	4	0	0	0	0	1	2
Level 5	3	0	0	1	2	0	0	0	0	0	1

Subject 9											
Level 1	4	4	1	4	3	0	0	1	0	2	3
Level 2	4	4	3	4	4	0	1	1	0	1	3
Level 3	4	2	2	4	3	1	1	0	1	1	3
Level 4	4	2	2	4	4	0	0	0	0	2	3
Level 5	4	3	2	3	4	0	0	2	1	2	3

Subject 10											
Level 1	4	2	2	0	2	0	0	0	1	1	3
Level 2	4	1	2	1	3	0	0	0	0	1	3
Level 3	4	1	2	2	2	0	0	0	1	1	2
Level 4	4	1	2	1	2	0	0	0	1	1	3
Level 5	3	2	1	1	2	0	0	0	0	1	2

Subject 11											
Level 1	4	2	2	2	3	0	0	1	1	3	3
Level 2	4	2	1	3	4	0	0	0	2	2	3
Level 3	4	1	2	3	4	0	0	0	1	1	2
Level 4	3	1	1	2	3	0	0	1	1	2	2
Level 5	1	1	0	1	2	0	1	1	1	1	1

A.4.5.2 Comparison Listening Effort Test 1 Rep. 2

Subject 1											
Level 1	4	1	3	2	4	0	1	1	2	2	4
Level 2	3	1	4	2	4	0	0	1	1	2	4
Level 3	4	2	1	1	4	0	0	1	2	1	4
Level 4	3	1	1	2	3	0	1	0	1	2	3
Level 5	2	1	0	1	3	0	0	0	2	1	3

Subject 2											
Level 1	4	1	2	1	4	4	1	1	2	3	3
Level 2	4	1	1	1	4	0	1	1	1	3	3
Level 3	4	1	1	2	3	0	1	0	1	3	2
Level 4	3	1	1	1	3	0	1	1	0	2	2
Level 5	3	2	0	0	3	1	0	0	1	2	3

Subject 3											
Level 1	4	2	1	3	3	0	0	0	0	3	2
Level 2	4	2	2	3	2	0	0	0	0	3	1
Level 3	4	1	2	2	3	0	0	0	1	3	2
Level 4	3	1	2	3	3	0	0	0	0	3	1
Level 5	3	1	1	2	2	0	0	0	1	2	1

Subject 4											
Level 1	4	2	3	2	3	0	0	2	1	2	3
Level 2	4	0	0	3	4	4	0	0	1	1	2
Level 3	3	0	3	2	3	0	0	0	0	1	4
Level 4	4	3	1	1	4	0	0	1	2	2	4
Level 5	3	2	1	1	3	0	1	2	1	1	4

Subject 5											
Level 1	3	1	1	2	2	0	1	0	1	1	3
Level 2	3	1	1	1	3	0	0	1	1	1	3
Level 3	3	2	0	1	4	0	1	0	1	2	2
Level 4	3	2	1	1	3	1	0	0	1	0	2
Level 5	2	0	0	0	2	0	1	0	0	1	2
Subject 6											
Level 1	4	3	0	3	4	0	0	2	1	3	3
Level 2	4	2	0	4	4	0	0	1	1	3	4
Level 3	4	1	1	2	4	0	0	0	0	2	2
Level 4	4	0	1	1	3	0	0	1	0	2	2
Level 5	3	1	0	1	2	1	0	2	0	0	1
Subject 7											
Level 1	3	3	2	1	3	0	0	0	2	2	2
Level 2	4	1	1	1	4	0	0	0	0	1	1
Level 3	4	2	1	2	4	0	0	0	1	1	2
Level 4	3	0	1	0	4	0	0	0	0	0	3
Level 5	4	1	1	0	3	0	0	0	0	1	0
Subject 8											
Level 1	4	1	1	2	4	0	1	0	0	1	2
Level 2	4	1	1	1	3	0	0	1	0	0	1
Level 3	3	1	0	0	3	0	0	0	0	1	2
Level 4	3	3	0	1	3	0	0	0	0	1	1
Level 5	2	0	1	1	2	0	0	0	0	0	2
Subject 9											
Level 1	4	2	1	3	4	0	1	2	1	3	3
Level 2	4	1	1	3	4	0	0	1	1	2	4
Level 3	4	1	1	3	4	1	0	0	0	2	4
Level 4	4	1	2	3	4	0	1	0	1	2	4
Level 5	4	0	1	4	4	1	0	1	1	2	4
Subject 10											
Level 1	4	1	2	1	1	0	0	0	0	0	2
Level 2	4	2	2	2	1	0	0	0	1	1	1
Level 3	4	0	1	1	2	0	0	0	1	1	1
Level 4	4	1	0	2	1	0	0	0	0	2	2
Level 5	3	1	0	1	1	0	0	0	0	1	2
Subject 11											
Level 1	4	0	2	1	2	0	0	0	0	0	4
Level 2	4	4	1	4	3	0	0	0	0	0	2
Level 3	4	1	0	2	3	0	0	0	0	1	1
Level 4	3	2	1	2	0	0	0	0	0	1	1
Level 5	3	0	0	2	1	0	0	0	0	0	2

A.4.5.3 Comparison Listening Effort Test 2 Rep. 1

Subject 1											
Level 1	4	2	3	2	4	4	3	4	3	2	1
Level 2	4	2	2	2	4	3	4	3	2	1	1
Level 3	4	2	2	2	4	4	3	3	2	2	1
Level 4	4	2	2	3	4	3	4	4	2	4	2
Level 5	4	3	1	3	4	3	4	3	2	2	0

Subject 2											
Level 1	4	0	1	1	4	4	3	4	2	0	0
Level 2	4	1	1	2	4	4	3	3	1	0	0
Level 3	4	0	1	1	4	4	3	2	1	0	0
Level 4	4	0	0	0	4	4	4	2	0	0	0
Level 5	4	0	0	0	4	3	3	2	1	0	0
Subject 3											
Level 1	4	1	0	3	4	4	4	3	2	1	0
Level 2	4	2	2	2	4	4	3	4	3	2	1
Level 3	4	1	1	3	4	4	4	3	3	2	0
Level 4	4	0	1	1	3	3	3	3	3	2	1
Level 5	4	0	0	1	2	3	3	2	2	1	0
Subject 4											
Level 1	3	1	2	1	4	3	3	2	2	0	0
Level 2	4	2	0	2	4	3	2	3	2	1	0
Level 3	4	1	1	0	4	4	2	2	1	1	0
Level 4	4	1	0	1	4	3	3	2	2	0	0
Level 5	3	1	0	1	4	4	2	2	2	1	0
Subject 5											
Level 1	4	1	1	0	4	4	4	3	2	1	1
Level 2	4	1	1	1	4	4	3	3	1	1	0
Level 3	4	2	2	2	3	4	3	2	1	1	1
Level 4	3	2	1	1	3	3	2	1	1	1	0
Level 5	3	1	1	1	2	2	2	2	1	1	0
Subject 6											
Level 1	4	1	0	1	4	3	4	3	1	1	0
Level 2	4	1	1	0	4	4	3	4	1	1	0
Level 3	4	0	0	1	4	3	3	3	1	0	0
Level 4	4	0	1	0	4	3	3	1	1	1	0
Level 5	4	1	0	1	4	3	3	2	0	1	0
Subject 7											
Level 1	4	4	2	2	4	4	3	3	3	2	0
Level 2	4	2	3	2	4	4	3	2	3	1	0
Level 3	4	2	1	2	4	4	4	4	2	1	0
Level 4	4	1	2	2	4	4	4	4	1	1	0
Level 5	3	1	1	1	4	3	3	3	2	1	0
Subject 8											
Level 1	4	0	1	0	4	4	4	3	2	1	0
Level 2	4	0	1	0	4	4	4	3	1	0	0
Level 3	3	2	0	0	4	2	4	2	1	0	0
Level 4	3	1	0	0	2	1	3	1	1	0	0
Level 5	2	0	1	0	1	1	2	1	0	0	0
Subject 9											
Level 1	4	1	1	1	4	4	3	2	2	0	0
Level 2	4	1	2	1	4	4	4	3	2	1	0
Level 3	4	1	1	1	4	4	3	3	2	0	0
Level 4	4	1	1	1	4	3	3	2	1	0	0
Level 5	4	0	1	1	3	3	3	2	2	0	0

Subject 10											
Level 1	4	0	2	1	3	1	3	2	1	1	0
Level 2	4	2	2	1	3	2	4	3	1	0	0
Level 3	4	0	1	0	3	2	2	2	1	1	0
Level 4	4	1	2	1	3	2	3	2	1	0	0
Level 5	4	2	0	1	4	1	3	3	1	0	0
Subject 11											
Level 1	4	1	2	1	4	4	4	4	3	2	0
Level 2	4	2	2	1	4	4	4	3	2	1	0
Level 3	4	2	1	0	4	4	4	3	4	1	0
Level 4	4	4	3	2	4	2	3	4	3	2	0
Level 5	3	0	0	0	4	2	3	4	1	0	1

A.4.5.4 Comparison Listening Effort Test 2 Rep. 2

Subject 1											
Level 1	3	1	2	1	4	4	2	3	1	1	1
Level 2	3	0	1	1	4	4	2	1	0	0	0
Level 3	4	1	1	1	4	3	3	2	1	0	0
Level 4	2	1	0	1	4	3	2	1	1	0	0
Level 5	1	2	0	1	3	3	3	2	0	1	0
Subject 2											
Level 1	4	1	3	2	4	4	4	3	3	2	1
Level 2	4	2	2	2	4	4	3	3	2	1	1
Level 3	4	1	2	2	3	4	2	3	2	2	0
Level 4	4	1	2	1	3	2	2	2	1	1	0
Level 5	3	1	1	1	4	3	4	3	1	1	0
Subject 3											
Level 1	4	1	0	2	2	4	3	2	1	0	0
Level 2	4	1	1	1	3	4	3	2	2	0	0
Level 3	4	1	1	1	2	4	2	2	1	1	0
Level 4	3	1	1	2	4	3	2	2	2	1	0
Level 5	3	1	0	0	2	3	3	1	1	0	0
Subject 4											
Level 1	2	1	3	2	2	3	3	3	3	0	0
Level 2	2	0	2	1	3	4	3	3	0	0	0
Level 3	3	0	1	1	3	3	3	3	3	0	1
Level 4	2	2	2	0	3	3	2	4	0	1	0
Level 5	1	1	0	1	3	3	3	4	2	0	0
Subject 5											
Level 1	4	2	3	1	4	4	4	4	1	1	1
Level 2	3	1	1	1	0	4	4	3	2	0	0
Level 3	4	2	2	2	4	4	4	4	2	1	0
Level 4	4	2	1	1	4	4	4	3	0	1	0
Level 5	3	1	0	1	3	3	3	4	2	0	0
Subject 6											
Level 1	4	2	1	2	1	4	4	2	3	1	0
Level 2	3	1	1	1	0	4	4	3	2	0	0
Level 3	4	2	2	1	0	4	4	3	1	0	0
Level 4	4	1	2	1	1	4	3	2	1	1	0
Level 5	3	1	1	1	1	3	3	3	0	1	0

Subject 7											
Level 1	3	4	2	2	4	4	3	3	1	2	0
Level 2	4	2	2	2	4	4	3	4	3	1	0
Level 3	4	3	1	2	4	4	4	3	2	2	0
Level 4	3	1	2	2	4	4	4	4	1	0	0
Level 5	1	1	0	1	4	4	4	2	1	0	0
Subject 8											
Level 1	3	1	2	0	4	3	3	4	2	1	0
Level 2	4	0	2	0	4	4	4	2	1	1	0
Level 3	4	1	1	0	3	4	4	3	1	1	0
Level 4	3	0	0	0	3	3	3	4	1	1	0
Level 5	2	0	0	0	1	2	1	1	1	0	0
Subject 9											
Level 1	4	2	1	1	4	2	3	2	1	0	1
Level 2	4	1	2	1	4	3	3	2	1	1	0
Level 3	3	3	1	1	4	1	3	1	1	0	0
Level 4	4	2	1	1	4	2	2	1	1	0	1
Level 5	3	1	0	0	4	1	3	1	0	1	0
Subject 10											
Level 1	4	1	1	2	4	2	3	3	1	2	0
Level 2	4	2	2	1	4	1	3	3	2	1	0
Level 3	4	2	2	0	4	2	2	2	2	1	0
Level 4	3	1	2	1	4	1	2	2	2	1	0
Level 5	3	2	1	1	3	1	2	1	1	0	0
Subject 11											
Level 1	4	0	1	1	4	3	4	3	1	1	0
Level 2	4	0	1	1	3	3	4	2	1	1	0
Level 3	4	1	2	1	3	2	4	2	1	0	0
Level 4	4	2	1	2	4	2	4	2	1	0	0
Level 5	4	1	0	1	4	2	4	2	1	0	0

A.4.5.5 Comparison Listening Effort Test 3 Rep. 1

Subject 1											
Level 1	2	0	0	0	4	1	4	2	0	0	1
Level 2	4	2	1	2	3	3	3	2	0	1	2
Level 3	4	2	0	1	4	4	4	3	1	0	2
Level 4	3	1	1	1	4	3	4	2	0	0	1
Level 5	3	1	0	0	4	2	4	3	1	0	1
Subject 2											
Level 1	4	1	3	3	4	3	2	2	2	1	2
Level 2	4	1	1	2	4	2	4	4	4	0	2
Level 3	4	0	1	3	4	3	2	3	2	0	3
Level 4	3	2	1	3	3	2	4	2	2	0	2
Level 5	2	0	0	2	2	2	2	1	1	2	3
Subject 3											
Level 1	4	0	3	0	3	3	3	4	1	0	0
Level 2	4	2	3	1	4	4	1	4	0	0	1
Level 3	4	0	3	1	4	4	2	3	0	0	0
Level 4	4	1	1	0	4	4	1	2	0	0	0
Level 5	4	1	0	0	4	3	2	1	0	0	1

Subject 4											
Level 1	3	1	1	2	4	3	3	3	3	0	2
Level 2	4	1	1	1	4	3	4	4	1	1	1
Level 3	4	1	1	0	4	3	4	4	1	0	1
Level 4	4	2	1	1	4	4	4	4	1	0	1
Level 5	4	1	0	1	4	3	3	3	1	0	1
Subject 5											
Level 1	4	1	2	2	4	4	2	3	1	1	2
Level 2	4	2	1	2	3	3	3	3	2	0	2
Level 3	4	2	2	2	4	3	3	2	1	1	0
Level 4	4	1	1	2	4	4	3	2	0	1	1
Level 5	4	2	0	2	4	3	3	2	1	0	2
Subject 6											
Level 1	4	2	1	2	4	4	4	4	3	1	2
Level 2	3	3	1	1	4	4	4	2	2	0	1
Level 3	3	2	0	2	4	3	4	1	1	0	0
Level 4	4	2	0	2	4	3	4	3	2	0	0
Level 5	4	0	0	3	4	3	4	3	1	0	0
Subject 7											
Level 1	4	2	1	2	3	4	2	1	2	0	2
Level 2	4	3	2	3	3	4	3	3	1	0	1
Level 3	3	2	1	1	4	3	2	2	1	1	2
Level 4	4	0	1	1	4	3	3	3	0	0	1
Level 5	3	3	1	1	3	3	3	2	2	1	2
Subject 8											
Level 1	3	2	3	1	3	3	3	1	1	0	0
Level 2	3	1	2	0	3	4	4	3	1	1	2
Level 3	4	2	1	2	3	4	4	4	1	0	0
Level 4	3	2	2	1	2	3	3	1	1	1	1
Level 5	3	0	0	0	1	2	3	1	0	0	0
Subject 9											
Level 1	4	1	0	0	3	2	4	2	1	0	1
Level 2	4	3	1	1	4	3	4	3	1	2	0
Level 3	4	2	1	2	4	3	4	2	1	0	1
Level 4	3	2	1	2	4	3	3	2	0	0	1
Level 5	3	1	1	0	3	2	4	1	1	0	0
Subject 10											
Level 1	4	2	2	3	4	2	4	2	1	2	1
Level 2	4	2	2	2	4	3	4	3	1	1	1
Level 3	4	2	2	1	4	2	4	3	2	0	0
Level 4	3	2	2	2	3	2	4	2	1	0	0
Level 5	3	1	2	0	3	2	3	3	0	1	0
Subject 11											
Level 1	4	4	3	2	3	4	4	4	3	3	1
Level 2	4	3	2	3	4	1	4	4	3	2	3
Level 3	4	3	4	3	4	4	4	4	1	3	0
Level 4	4	2	1	4	4	4	4	4	0	2	3
Level 5	4	3	2	3	4	3	4	4	3	2	0

A.4.5.6 Comparison Listening Effort Test 3 Rep. 2

Subject 1											
Level 1	4	2	1	2	4	4	4	4	1	2	0
Level 2	4	2	0	2	4	4	4	4	2	1	2
Level 3	4	2	2	2	4	4	4	4	2	1	3
Level 4	4	3	3	3	4	4	4	4	3	1	2
Level 5	3	1	1	2	4	4	3	4	2	0	1
Subject 2											
Level 1	4	1	0	1	2	2	2	2	2	0	0
Level 2	4	1	1	1	3	3	2	2	1	0	0
Level 3	4	1	0	1	3	3	3	0	1	0	0
Level 4	4	1	0	1	2	3	2	1	1	0	0
Level 5	3	0	0	1	2	2	3	1	1	0	0
Subject 3											
Level 1	4	1	1	0	4	4	2	3	2	1	1
Level 2	4	1	1	2	3	4	2	1	2	0	1
Level 3	4	1	3	2	4	3	3	1	2	0	0
Level 4	3	1	1	2	3	3	3	2	1	1	0
Level 5	2	2	4	1	3	2	2	1	1	0	1
Subject 4											
Level 1	4	1	2	3	4	4	4	3	1	1	1
Level 2	4	2	3	3	4	3	4	4	1	0	1
Level 3	4	3	3	2	4	4	4	4	3	0	2
Level 4	3	2	2	3	3	4	3	4	2	0	1
Level 5	3	1	1	2	3	4	4	4	1	2	2
Subject 5											
Level 1	4	2	1	3	4	4	4	2	1	2	1
Level 2	4	4	1	4	4	4	4	4	1	1	3
Level 3	4	3	1	4	4	4	3	4	4	2	3
Level 4	4	3	3	4	4	4	4	4	3	1	3
Level 5	4	4	0	2	4	4	4	4	0	3	3
Subject 6											
Level 1	4	2	0	2	4	3	3	3	1	0	2
Level 2	4	1	2	1	4	3	4	2	0	0	1
Level 3	3	1	0	3	3	4	3	3	0	1	0
Level 4	4	2	1	3	4	3	1	3	1	1	0
Level 5	4	1	1	2	3	3	4	2	0	0	1
Subject 7											
Level 1	4	2	3	2	3	4	4	4	3	0	3
Level 2	4	4	1	0	4	4	4	4	2	1	2
Level 3	4	4	2	3	4	4	3	3	1	0	1
Level 4	3	3	2	4	3	4	3	2	1	0	1
Level 5	3	3	1	2	3	3	3	2	0	1	1
Subject 8											
Level 1	4	4	3	0	3	2	3	2	2	1	1
Level 2	3	3	2	2	4	4	4	4	3	1	2
Level 3	4	3	3	2	4	4	4	2	0	2	1
Level 4	4	2	0	1	4	3	4	3	1	0	1
Level 5	4	1	1	1	3	3	3	1	2	1	0

Subject 9											
Level 1	4	2	1	1	4	4	3	3	1	1	1
Level 2	4	2	1	1	4	4	3	4	2	0	1
Level 3	4	3	1	1	3	2	4	3	2	0	1
Level 4	4	3	1	2	4	4	4	4	1	1	1
Level 5	4	3	1	1	4	3	4	2	2	0	0
Subject 10											
Level 1	4	2	2	3	4	3	2	3	2	2	1
Level 2	4	2	1	2	4	2	3	2	1	1	0
Level 3	4	2	1	2	4	3	3	2	2	1	1
Level 4	4	2	2	2	4	2	3	3	1	1	0
Level 5	4	1	2	2	4	2	3	2	2	1	1
Subject 11											
Level 1	3	2	1	1	3	1	4	4	1	0	1
Level 2	4	2	2	2	4	3	3	4	2	0	1
Level 3	4	1	2	3	4	3	4	3	2	0	0
Level 4	4	1	0	2	3	3	4	2	2	0	1
Level 5	4	1	3	0	4	4	4	3	1	0	0

A.4.5.7 Comparison Listening Effort Test 4 Rep. 1

Subject 1											
Level 1	4	2	1	3	4	4	3	2	3	0	1
Level 2	4	1	2	3	4	3	3	3	1	0	1
Level 3	4	1	2	3	4	3	3	3	0	0	1
Level 4	4	2	0	3	3	3	3	3	1	0	0
Level 5	4	1	0	2	3	4	2	2	1	0	0
Subject 2											
Level 1	4	2	0	3	4	4	4	2	0	0	3
Level 2	4	3	2	3	4	4	3	3	3	0	0
Level 3	4	2	1	3	4	4	4	3	2	0	1
Level 4	4	2	1	3	4	4	4	2	0	0	0
Level 5	4	0	0	3	4	3	4	2	0	0	0
Subject 3											
Level 1	4	1	2	3	3	2	2	2	3	1	1
Level 2	3	2	1	3	3	4	2	2	3	1	0
Level 3	4	2	1	4	3	3	3	2	2	0	1
Level 4	4	2	1	3	2	4	3	2	1	0	1
Level 5	3	1	0	3	2	3	4	1	2	1	0
Subject 4											
Level 1	4	1	1	2	4	2	2	3	1	0	0
Level 2	4	2	1	2	3	3	2	1	1	0	0
Level 3	3	1	2	3	4	3	2	3	1	1	1
Level 4	3	1	1	2	4	3	2	2	2	0	1
Level 5	3	1	1	2	3	2	2	1	1	0	0
Subject 5											
Level 1	4	2	1	3	3	4	3	1	0	0	0
Level 2	4	3	0	3	3	3	2	2	1	2	2
Level 3	4	3	2	3	3	4	3	1	1	1	0
Level 4	4	3	0	2	3	3	3	2	1	0	2
Level 5	4	2	2	3	3	3	3	1	0	1	1

Subject 6											
Level 1	4	2	1	2	4	4	4	4	2	0	4
Level 2	4	2	0	4	4	4	4	4	3	1	2
Level 3	4	1	1	3	4	4	4	3	3	1	0
Level 4	3	1	1	2	3	3	3	2	1	0	1
Level 5	2	1	0	2	2	2	2	1	2	0	1

Subject 7											
Level 1	4	1	2	3	4	4	2	3	3	3	2
Level 2	4	1	2	3	4	4	3	3	2	3	2
Level 3	4	1	3	4	3	4	4	3	2	2	1
Level 4	3	2	2	1	4	4	4	2	2	1	1
Level 5	3	2	2	1	3	3	2	2	2	0	1

Subject 8											
Level 1	4	2	1	3	4	3	4	3	2	1	2
Level 2	4	3	2	2	3	4	4	3	2	2	1
Level 3	4	2	2	2	4	4	3	2	2	1	3
Level 4	4	1	1	3	4	4	4	3	3	1	1
Level 5	4	1	1	2	4	3	3	3	2	0	2

Subject 9											
Level 1	4	1	1	3	4	3	3	2	0	0	0
Level 2	4	1	2	3	4	4	4	3	1	1	1
Level 3	4	1	2	3	4	4	3	3	0	0	1
Level 4	4	2	2	3	4	3	4	2	0	0	1
Level 5	4	2	1	3	4	4	4	2	1	1	1

Subject 10											
Level 1	3	2	2	3	3	2	3	2	1	0	2
Level 2	3	2	1	3	4	0	1	3	2	1	0
Level 3	4	1	2	4	4	1	4	3	2	0	1
Level 4	4	1	1	3	4	1	2	2	1	0	1
Level 5	4	1	0	2	3	1	3	1	1	0	0

Subject 11											
Level 1	4	4	2	4	4	4	4	3	2	1	3
Level 2	4	3	1	4	4	4	4	4	3	2	1
Level 3	4	3	2	4	4	4	4	2	1	1	3
Level 4	4	0	4	4	4	4	4	4	2	2	0
Level 5	4	0	1	4	4	4	4	2	3	2	0

A.4.5.8 Comparison Listening Effort Test 4 Rep. 2

Subject 1											
Level 1	4	2	2	2	4	4	4	3	2	2	2
Level 2	4	2	2	2	4	4	4	3	1	0	1
Level 3	4	2	1	3	4	4	4	3	0	1	2
Level 4	4	2	1	2	4	4	4	3	1	0	2
Level 5	3	3	1	1	4	4	4	3	2	1	1

Subject 2											
Level 1	4	2	1	3	3	3	2	2	2	1	1
Level 2	4	1	2	3	3	3	2	2	1	1	1
Level 3	3	0	1	3	3	3	2	2	1	0	0
Level 4	3	0	0	2	3	3	3	2	1	0	0
Level 5	3	0	1	1	3	3	2	1	0	0	0

Subject 3											
Level 1	4	0	0	2	1	1	1	3	1	0	0
Level 2	4	1	0	1	3	2	1	2	1	0	1
Level 3	3	1	0	2	3	2	2	2	2	0	0
Level 4	3	1	0	1	2	3	2	2	2	0	0
Level 5	4	1	1	1	3	2	1	2	2	1	1

Subject 4											
Level 1	4	3	2	4	4	4	3	4	2	0	2
Level 2	4	1	2	3	4	4	4	3	3	0	2
Level 3	4	3	2	4	4	4	3	4	3	0	3
Level 4	4	2	2	4	4	4	4	4	3	1	3
Level 5	4	3	1	3	4	4	4	4	2	2	2

Subject 5											
Level 1	4	3	3	4	4	4	3	3	1	2	4
Level 2	4	4	3	4	4	4	4	4	2	2	4
Level 3	4	3	3	4	4	4	4	4	2	1	3
Level 4	3	2	2	4	4	4	3	3	1	1	3
Level 5	4	2	1	3	3	4	3	3	0	1	2

Subject 6											
Level 1	4	2	0	3	4	4	4	3	2	1	1
Level 2	4	2	1	4	4	4	4	4	2	1	1
Level 3	4	2	1	4	4	4	4	4	2	2	1
Level 4	4	2	1	4	4	4	3	3	3	0	0
Level 5	4	1	0	3	4	3	3	3	0	1	3

Subject 7											
Level 1	4	1	1	1	3	4	3	1	1	0	1
Level 2	4	1	1	1	4	4	3	2	2	0	2
Level 3	4	1	1	1	4	4	3	1	1	1	1
Level 4	4	0	1	2	4	4	2	3	1	0	0
Level 5	4	1	0	2	4	4	1	3	2	1	0

Subject 8											
Level 1	3	2	3	2	3	3	3	1	1	0	0
Level 2	3	1	2	1	3	4	4	3	1	1	2
Level 3	4	2	1	3	3	4	4	4	1	0	0
Level 4	3	2	2	2	2	3	3	1	1	1	1
Level 5	3	0	0	2	1	2	3	1	0	0	0

Subject 9											
Level 1	3	2	1	2	4	3	3	3	1	1	1
Level 2	4	2	2	3	4	4	3	3	2	0	1
Level 3	3	1	2	3	3	3	3	3	1	0	1
Level 4	3	1	1	2	2	3	3	2	1	0	1
Level 5	2	0	0	2	3	2	2	2	1	0	0

Subject 10											
Level 1	4	2	3	4	4	3	3	3	1	1	2
Level 2	3	3	2	3	3	3	3	2	3	1	2
Level 3	4	2	2	3	4	3	3	3	2	0	1
Level 4	3	2	3	2	4	2	3	3	2	1	1
Level 5	3	3	3	3	3	2	3	2	1	0	2

Subject 11

Level 1	4	2	1	2	4	2	4	4	2	2	1
Level 2	4	2	4	4	3	1	4	3	1	2	2
Level 3	3	1	2	3	4	2	4	3	1	1	0
Level 4	4	1	3	3	4	2	3	1	1	1	1
Level 5	3	1	2	2	4	1	3	2	2	0	0

A.4.6.1 MNRU Listening Effort Test Rep.1

Subject 1

Level 1	4	4	1	4	4	4	4	4	0	0	0
Level 2	4	4	1	4	4	4	4	4	0	1	2
Level 3	4	4	1	3	4	4	4	4	0	0	0
Level 4	4	3	1	3	3	3	3	3	0	0	0
Level 5	4	3	0	2	3	3	3	2	0	0	1

Subject 2

Level 1	3	4	1	2	4	2	4	3	1	1	2
Level 2	4	4	2	3	4	3	4	3	2	2	2
Level 3	4	4	1	3	4	3	4	3	3	0	1
Level 4	4	4	1	3	4	3	4	3	2	1	0
Level 5	4	4	1	2	4	4	4	3	2	0	2

Subject 3

Level 1	2	3	0	3	3	2	4	1	1	0	1
Level 2	2	2	0	2	2	3	4	2	1	1	0
Level 3	3	3	0	1	3	2	4	2	1	0	1
Level 4	2	2	1	3	2	2	3	3	2	1	1
Level 5	2	2	0	2	2	2	4	1	1	0	0

Subject 4

Level 1	4	3	2	3	4	4	4	3	1	0	0
Level 2	4	3	2	3	4	4	3	3	1	1	1
Level 3	4	3	2	3	4	3	4	3	1	0	1
Level 4	4	2	2	3	4	3	4	3	1	1	1
Level 5	3	3	2	3	4	3	3	2	1	0.	1

Subject 5

Level 1	4	4	1	3	4	4	4	4	2	1	2
Level 2	4	4	1	4	4	3	4	2	2	1	3
Level 3	4	4	2	4	4	4	4	3	2	0	2
Level 4	4	4	2	4	3	3	4	4	2	1	2
Level 5	4	4	1	3	3	3	3	3	1	1	3

Subject 6

Level 1	4	4	2	4	4	4	4	3	1	2	1
Level 2	4	4	3	3	4	4	4	4	2	1	1
Level 3	4	4	2	3	4	4	4	4	0	0	1
Level 4	4	3	2	4	4	4	4	3	0	0	0
Level 5	3	3	3	3	4	4	4	4	1	0	2

Subject 7

Level 1	4	4	1	3	3	4	2	1	1	0	0
Level 2	4	4	1	2	4	4	3	2	0	0	0
Level 3	4	4	1	3	3	3	4	2	1	2	1
Level 4	4	3	2	3	4	4	3	2	0	1	1
Level 5	4	2	2	1	3	3	3	1	0	2	1

Subject 8											
Level 1	4	3	2	4	4	3	4	4	3	3	3
Level 2	4	3	2	4	4	4	4	4	3	3	3
Level 3	4	3	2	3	4	3	4	3	3	2	3
Level 4	4	3	2	4	4	4	4	4	2	3	3
Level 5	4	3	2	3	3	4	4	3	2	3	3
Subject 9											
Level 1	4	3	1	2	4	2	3	1	0	1	1
Level 2	3	3	0	3	4	3	3	2	0	0	1
Level 3	4	4	2	3	4	3	4	1	1	0	2
Level 4	3	4	1	2	4	2	4	2	1	1	2
Level 5	3	3	1	2	3	2	3	2	1	1	1
Subject 10											
Level 1	4	4	1	2	4	1	4	4	3	1	1
Level 2	4	4	2	3	4	2	3	3	3	2	1
Level 3	4	4	2	3	4	2	4	3	1	0	1
Level 4	4	3	2	3	4	1	4	3	2	2	0
Level 5	4	4	1	2	4	1	4	1	1	0	2
Subject 11											
Level 1	4	4	1	3	4	2	4	3	2	1	1
Level 2	4	4	1	3	4	1	4	4	2	2	2
Level 3	4	4	2	3	4	2	4	4	1	1	1
Level 4	4	4	2	3	4	2	3	3	2	1	1
Level 5	3	3	1	3	4	2	3	3	1	1	1

A.4.6.2 MNRU Listening Effort Test Rep.2

Subject 1											
Level 1	4	3	2	2	4	4	2	1	2	0	0
Level 2	3	4	1	3	4	4	4	4	2	1	0
Level 3	4	4	3	1	3	4	3	4	2	0	1
Level 4	3	2	3	2	2	2	2	3	1	0	0
Level 5	2	3	1	2	3	3	2	2	2	1	1
Subject 2											
Level 1	4	3	1	2	3	2	2	1	1	1	1
Level 2	3	3	0	3	4	2	2	1	1	1	1
Level 3	4	2	1	2	3	3	2	2	1	0	1
Level 4	4	2	1	3	4	3	3	2	2	0	0
Level 5	4	2	2	3	3	2	3	1	0	0	1
Subject 3											
Level 1	3	2	0	2	3	3	3	1	2	0	1
Level 2	4	3	0	2	4	3	4	2	2	0	1
Level 3	4	3	0	2	3	3	4	1	2	0	1
Level 4	3	3	1	2	2	3	4	2	2	0	1
Level 5	3	3	1	2	2	3	3	2	1	1	0

Subject 4

Level 1	4	3	1	2	3	3	3	3	1	1	1
Level 2	4	3	2	3	4	3	3	2	1	3	1
Level 3	4	3	2	3	4	4	4	2	1	0	1
Level 4	4	3	2	3	4	4	3	2	2	0	1
Level 5	3	4	1	3	4	3	3	2	1	0	0

Subject 5

Level 1	4	4	3	4	4	4	4	2	1	1	1
Level 2	4	4	2	4	4	4	4	2	1	0	2
Level 3	4	4	3	3	4	4	4	3	0	0	1
Level 4	4	4	3	4	4	4	4	3	0	1	2
Level 5	4	3	3	3	3	4	4	3	1	0	2

Subject 6

Level 1	4	3	1	2	3	2	3	3	1	1	0
Level 2	4	2	1	2	4	1	4	3	2	0	1
Level 3	4	2	1	1	4	1	3	3	0	0	1
Level 4	3	2	0	2	1	1	2	2	1	1	1
Level 5	3	2	0	1	1	1	2	2	1	0	1

Subject 7

Level 1	4	4	1	3	3	4	1	1	1	0	0
Level 2	4	4	1	2	4	4	3	2	0	0	1
Level 3	4	3	2	2	3	3	3	2	1	2	1
Level 4	4	3	2	3	4	4	3	1	0	1	1
Level 5	4	2	2	1	3	3	3	1	0	1	0

Subject 8

Level 1	4	3	1	2	4	2	3	1	0	1	1
Level 2	3	3	0	3	4	3	3	2	0	0	1
Level 3	4	4	2	3	4	3	4	1	1	0	2
Level 4	3	4	1	2	4	2	4	2	1	1	2
Level 5	3	3	1	2	3	2	3	2	1	1	1

Subject 9

Level 1	4	3	2	4	4	3	4	4	3	3	3
Level 2	4	3	2	4	4	4	4	4	3	3	3
Level 3	4	3	2	3	4	3	4	3	3	2	3
Level 4	4	3	2	4	4	4	4	4	2	3	3
Level 5	4	3	2	3	3	4	4	3	2	3	3

Subject 10

Level 1	4	4	1	4	4	4	4	4	0	0	0
Level 2	4	4	1	4	4	4	4	4	0	1	2
Level 3	4	4	1	3	4	4	3	4	0	0	0
Level 4	4	3	1	3	3	3	3	3	0	0	0
Level 5	4	3	0	2	3	3	3	2	0	0	1

Subject 11

Level 1	4	4	2	4	4	2	4	3	3	2	1
Level 2	4	4	3	3	4	3	4	4	2	2	2
Level 3	4	4	2	3	4	3	4	3	2	3	2
Level 4	3	4	1	4	4	3	3	3	3	2	1
Level 5	3	4	2	3	4	2	4	3	2	1	2

References

References

1- Allen Johnathon From Text to Speech: the MITalk System, Edited by Johnathan Allen, Cambridge University Press 1986

2- Atal, B. S. and Hanaauer, S. L. `Speech Analysis and Synthesis by Linear Prediction of the Speech Wave`, Journal of the Acoustic Society of America, vol.50, pp 637-655 (1971)

3- Bacri, N. `Perceptual Spaces and the Identification of Natural & Synthetic Sentences` Copy of proposed paper for CCITT.

4- Bristow, G. J. `Overview of Speech Output Devives.`, 1st Conf. on Speech Technology, Brighton 1984

5- Cooper, F.S., Liberman, A. M. and Borst, J.M., `Some Experiments on the Perception of Synthetic Speech Sounds ', In: Flanagan, J. L. & Rabiner, L. R. (Ed), Benchmark papers in acoustics, Speech Synthesis, Dowden, Hutinson and Ross.

6- CSELT, Italy, `Subjective assessment of automatic voice answering machines`. Source:- (CCITT COM XII - R The document is annexed to the reply to question 5/XIL.) 12, October 1986, pp 6-11.

7- Fairbanks, G, `Test of Phonetic Differentiation: The Rhyme Test, Journal of the Acoustic Society of America, vol.30, pp 596 (1958)

8- Greenspan, S.L., Nusbaum, H.C. & Pisoni, D. B. `Perception of Speech Generated by rule: Efects of Training and Attentional Limitation` Research on Speech Perception, Progress Report No. 11, Indiana University 1985

9- Huggins, A. W. F. (1964) `Distortion of the temporal pattern of speech: interruption and alternation`, Journal of the Acoustic Society of America, vol.36, pp 1055 - 64.

10- Johnston, R. D. ` Speech input / out assessment: the user's requirement ' Journal of the Acoustic Society of America, Presented at Voice Proceeing: Online Publications, Pinner, UK, 1986

11- Kruskal, J. B., Wish, M. , 'Multidimensional Scaling ', Sage Publications, (1983)

12- Lawrence, W., `The Synthesis of Speech from Signals which have a Low Information Rate`., Communication Theory edited by w. Jackson (Butterworth Scientific Publications, London 1953)

13- Liberman, A. D., Cooper, F. S., Shankweiler, D. P. and Studdert-Kennedy, M. (1967) `Perception of the speech code`, Psychological Review, vol. 74, pp 431-61.

14- Logan, J.S., Pisoni, D. B. & Greene, B.G.,`Measuring the Segmental Intelligibility of Synthetic speech: results from Eight text-to-Speech Systems`, Research on Speech Perception, Progress Report No. 11, Indiana University 1984

15- Luce,P.A.., `Comprehension of Fluent Synthetic Speech Produced by Rule` Research on Speech Perception, Progress Report No. 7, Indiana University 1981

16- Manous, L. M., & Pisoni, D. B., `Effects of Signal Duration on the Perception of Natural and Synthetic Speech` Research on Speech Perception, Progress Report No. 10, Indiana University 1984

17- Munson, W. A. & Karlin, J. E. ' Isopreference method for evaluating speech-transmission circuits '. Journal of the Acoustic Society of America, 1962 vol.34, pp 762-774.

18- McGee, V. E., `Determining Perceptual Spaces for the Quality of Filtered Speech`, Journal of Speech and Hearing Research, 1965, Vol 8, pp23-28

19- MalSheen, B. J., Wright,T. J., Yue, M. & Peet, M., `Intelligigibility of Synthetic CVC Stimuli over the Telephone`, Paper presented at the 111th meeting of the Acoustical Society of America, Cleveland, Ohio May 1986.

20- Miller, G. A. (1962) `Decision units in the perception of speech`, Transactions in Information Theory, vol 8, pp 81-83

21- Morrison, D. F.,'Multuivariate Statistical Methods', 2nd edition, McGraw-Hill, International Editions (1988), pp 266-302 (Principal Component Analysis)

22- Morrison, D. F.,'Multuivariate Statistical Methods', 2nd edition, McGraw-Hill, International Editions (1988), pp 95-97 (multiple regression)

23- Nusbaum, H.C. & Pisoni, D. B., `Constraints on the Perception of Synthetic speech. Generated by Rule`, Research on Speech Perception, Progress Report No.10, Indiana University 1984

24- Nusbaum, H.C., Schwab,C.E. & Pisoni, D. B. `Subjective Evaluation of Synthetic Speech, Measuring Preference, Naturalness and Acceptability` Research on Speech Perception, Progress Report No. 10, Indiana University 1984

25- Nye, P.W. & Gaitenby, J,H., `Consonant Intelligibility in Synthetic Speech and in Natural Speech Control (Modified Rhyme Test Results)`, Haskins Labs. SR-33, pp 77-91

26- Pisoni, D. B. `The Human Listener as a Cognitive Interface` Speech Technology, Vol 1, Number 2, April 1982

27- Pols, L.C.W. & Olive, J.P., `Intelligibility of Consonants in CVC Utterances Produced by Dyadic rule Synthesis`, Speech Comm. vol 2, pp3-13. 1983

28- Pratt, R. L., `Quantifying the Performance of Text-to-speech Synthesizers`, Copy of draft article for Speech Technology.

29- Richards, D. L., 'NOSFER Description and Reference Equivalent Proceedure', Telecommunication by Speech, pp 131-134, Butterworth &Co (Publishers) Ltd.

30- Richards, D. L., ' Modulated Noise Reference Unit ', Telecommunication by Speech, pp 295, Butterworth &Co (Publishers) Ltd.

31- Simpson, C.A., McCauley, M.E., Roland, E.F., Ruth, J.C., & Williges, B.H. `System Design for Speech Recognition and Generation`, Human Factors, 1985, vol 27, pp115-141

32- Slowiaczek, L.M. and Pisoni,D. B., `Effects of Practice on Speeded Classification of Natural and Synthetic Speech.` Research on Speech Perception, Progress Report No. 7, Indiana University 1981

33- Steinhieiser, F. H. and Burrows, D. J. (1973) `Chronometric analysis of speech perception`, Perception and Psychology, vol. 13.pp 426 - 30.

34- Sweden, `Subjective Quality Assessment of Synthetic Speech`, CCITT Working party XII/3, Question: 5/XII, Budapest, 4-6 May 1987

35- Voiers, W. D. (1977) ' Diagnostic accepability measure for speech communication systems ', Proc. IEEE- ICASSP77, pp 204-207

36- Voiers, W. D. (1977) ' Diagnostic evaluation of speech intelligibility ', In: M. Hawley (Ed), Benchmark papers in acoustics, Vol 11 , Speech intelligibility and speaker regognition, Dowden, Hutinson and Ross, Stroudsburg.

37- Voiers, W. D. (1982) ' Measurement of intrinsic deficiency in transmitted speech: The diagnostic Discrimination Test ', Proc. IEEE-ICASSP82, pp 1004-1007

38- Warren , R. M. and Obusek, O. J. (1971) `Speech perception and phonemic restorations`, Perception and Psychophysics, vol.9, pp 358 - 61.