A Phonetic Tactile Speech Listening System

CUED/F-INFENG/TR122

E.M. Ellis and A.J. Robinson

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, UK
Enquiries to: eme@uk.ac.cam.eng

May 1993

Abstract

Much of the current work on tactile aids concentrates on using the raw acoustic speech signal or close derivatives to stimulate the skin as supplementary information to aid in lip reading. The information rate per skin area associated with many of these schemes is too high to be processed accurately and thus limited in how well they can assist in speech listening. The scheme in this report presents speech information to the skin as phonetic symbols, which exhibits a lower information rate whilst maintaining a high proportion of the original speech information.

As a speech listening system a recurrent error propagation network phoneme recogniser will be the source of phonetic information to drive the tactile display. Phonetic information based on the standard DARPA TIMIT Acoustic Phonetic Continuous Speech Database has been reconstructed via a standard speech synthesiser (MITALK). The results are found to be reasonably intelligible showing that the phonetic format preserves a high percentage of the original speech information. The effects of pitch contours on the resulting speech quality is also investigated and the output from the recogniser is resynthesised to see how it compares with error-free phonetic information.

Suitable hardware has been designed for providing vibratory and pulse stimuli to the skin. Studies and experiments on the sensitivity of the skin show that the data rates associated with the phonetic format are similar to the rates that can be processed through the sense of touch.

To complete the scheme, a two-dimensional map of the phonemes of the English language has been formulated for the tactile display. The map is self-organising and exhibits spatial representation according to phoneme similarities.

INTRODUCTION

It is interesting to know how well one human sense can be made to act as a surrogate for one of the other senses, as is the case for tactile speech hearing aids. For tactile aids the skin is made to interpret some or all of the acoustic speech information that would normally be processed through the sense of hearing and many problems arise as a result.

Many single and multi-channel tactile devices have been developed to aid in the listening of speech e.g. [Weisenberger and Russell, 1989; Brookes and Frost, 1983], but these achieve limited levels of accuracy or are constrained to a very small vocabulary. One possible reason for the limited success is that the data rate per skin area associated with these schemes are too high to be processed with any clarity. At a receptor level, the skin is less sensitive than the ear by an estimated 14 orders of magnitude [Sherrick, 1985] rendering it incapable of processing complex stimulatory signals. By presenting the speech information to the skin as phonetic symbols, the information rate is suitably lowered to a level that can be more readily processed through the sense of touch.

There are three main problems associated with the phonetic tactile approach. Firstly, it is necessary to examine whether or not there is sufficient information in the phonetic format for intelligible tactile communication. Reducing the acoustic speech signal to its phonetic form is naturally accompanied by some loss in useful information. This loss should be large enough to suitably lower the tactile data rates, yet still retain the essential meaning in the original speech. The second problem area is linked with the first in that the data rate should be low enough to be conveyed by the standard stimulatory techniques that are available for this task. The final problem area is the way in which the individual phonemes are to be organised on a tactile display. The spatial representation must be natural from a perception view point, and have the ability to cope with errors in the phoneme data. The source of phonetic data to the tactile system is a recurrent error propagation phoneme recogniser [Robinson, 1992] and is the recogniser referred to throughout this report.

INFORMATION CONTENT OF PHONETIC FORMAT

It is accepted that there will be some loss of useful information if the input speech is reduced to a sequence of phonemes and then reconstructed again. Phonemes are the smallest linguistic units that can be used to distinguish meaning, and are thus the chosen units for this tactile system. It is important to investigate whether or not the phonetic format maintains enough of the information in the original speech for intelligible perception, or more importantly for tactile communication.

The recurrent net recogniser has been trained and tested on the DARPA TIMIT Acoustic Continuous Speech Database (hereafter referred to as the TIMIT database) which is a large vocabulary, multiple speaker database [Lamel *et al.*, 1987].

Information Degradation

There are a number of points along the chain, from the original utterance, to perception by the individual, where useful information can be lost. The TIMIT database is a comprehensive one made up of the original speech recorded as 16-bit 16KHz waveforms, as well as hand-labelled phonetic transcriptions. The waveform and labels are used in training and testing of the recurrent net recogniser, but are also useful for performing tests to determine the information content of the phonetic format when compared with the original speech. The various stages along the path from speech utterance to speech perception under the scheme described here, are shown in figure 1. By analysing the different paths it will be possible to determine the losses that can be tolerated at each stage, to arrive at an acceptable total loss through the system. Psycho-acoustic tests were performed which involved the perception of re-synthesised phoneme strings, but also included the perception of the original recorded speech for comparison.

Synthesis from Recognition

The process of re-synthesis from the phonetic format has been performed, and the results analysed for intelligibility. The sentences used in the tests that follow form over half of the core test set of the TIMIT database, which contains 192 sentences, spoken by 24 different speakers, from eight dialect regions in the united states. No limitations were put on the vocabulary or speaker variation and the criteria used for selecting sentences was the word count; any sentence containing more than nine words was considered too long to remember whilst writing down.

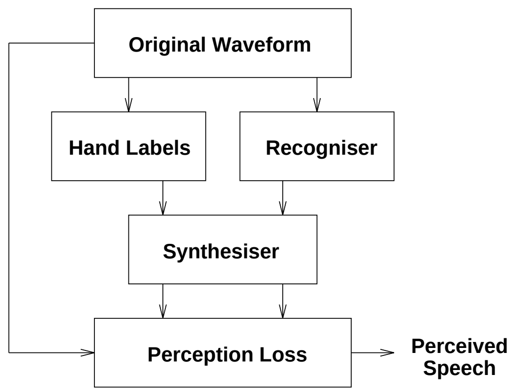

Figure 1: The Processing Paths used in Tests

With reference to figure 1, there are three possible sentence formats which will be used in the psycho-acoustic tests that follow:

(i) the original speech waveform
(ii) speech re-synthesised from the database hand labels
(iii) speech re-synthesised from the output of the phoneme recogniser

The synthesiser used in these tests is DECtalk DTC01, which is the hardware version of the MITalk speech-to-text system, based on the Klatt Synthesiser [Allen *et al.*, 1987]. DECtalk is able to accept phonemes (as well as text) and allows the pitch and duration to be specified for each phoneme.

Including Pitch Contours

It is of interest to know how intonation and stress can improve the perception of re-synthesised speech. If perception ability is greatly improved when intonation and stress is included, then this feature will have to be considered at a tactile level. The complete TIMIT data base has been pitch-tracked by [Tuerk, 1993] and has been used here to re-introduce the intonation and some components of the stress that was present in the original speech.

The pitch information has been recorded in frames of approximately 5ms which means that the pitch contours for each phoneme in the database is well detailed and sufficient for accurately re-introducing the pitch contours to the re-synthesised speech. The average phoneme length in the TIMIT database is approximately 80ms so there will be many frames of pitch information for each phoneme. The synthesiser will accept an optional argument of the target pitch for each phoneme it takes as input. DECtalk attempts to be at the specified pitch at the end of the relevant phoneme and uses its knowledge of phoneme length and type to give the appropriate pitch changes throughout phoneme. In doing this the synthesiser avoids monotonic phones which will make the speech sound disjointed and unnatural. Figure 2 shows how the frames from the pitch-tracker and the frames output from the recogniser come together to form input suitable for the synthesiser. It is interesting to note that the overall data rate of speech information to the synthesiser is increased only by a factor of about two (adding a value for the target pitch to each phone/phone duration) when the pitch information is added. This results in a very low bit-rate data stream suitable for transmission over slow or restricted communication channels. Experiments detailed later on in this report shows that the resulting re-synthesised speech is highly intelligible if there are few errors in the phoneme strings.

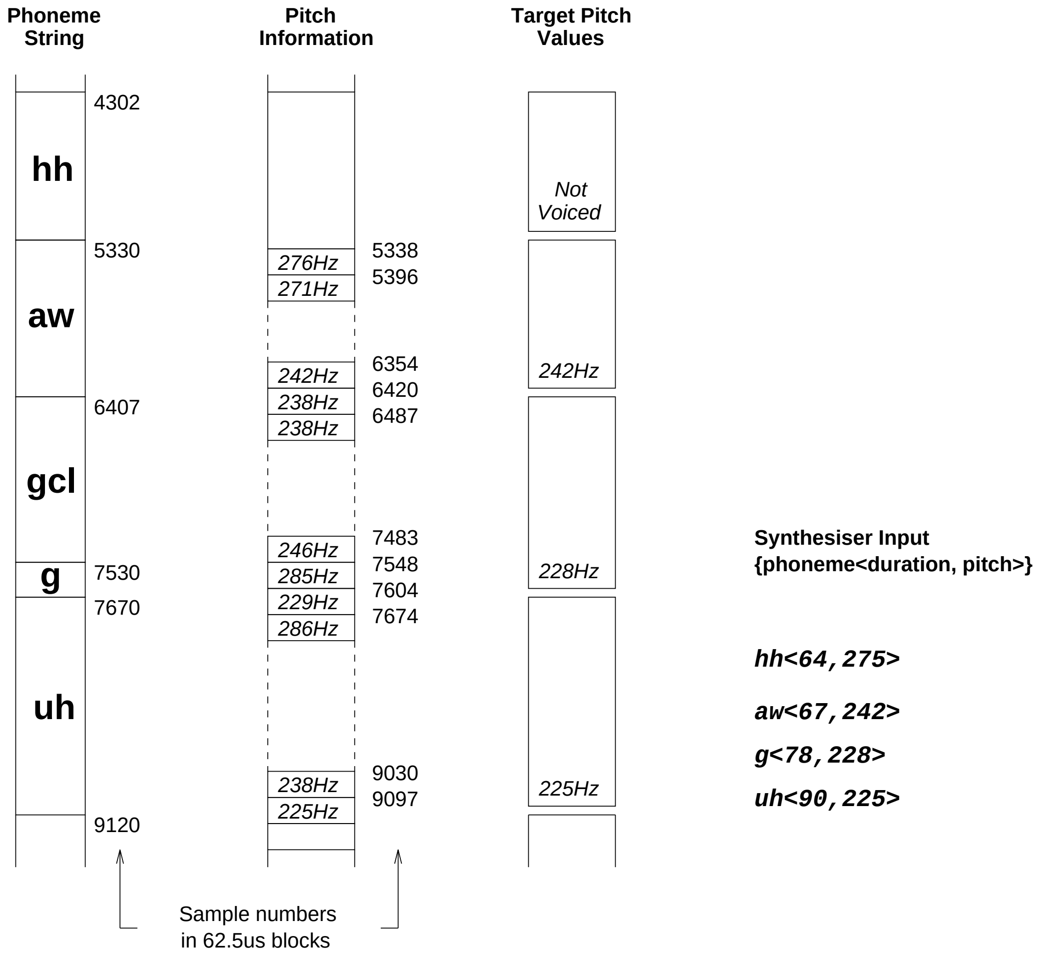

Figure 2: Phoneme and Pitch Frames for Re-synthesis

By adding the pitch contours to the phoneme sequences processed by DECtalk, two new sentence formats are formed:

(iv) pitch-tracked re-synthesised speech from the database hand labels
(v) pitch-tracked re-synthesised speech from the recogniser output

In total there are five different sentence formats of varying quality that are to be used in the experiments that follow.

Psycho-acoustic Tests

In the experiments performed, twelve subjects each listened to 45 sentences and were asked to write down what they heard. The sentences were a random mix of the five formats discussed in the previous section. Each of the five sentence formats were ultimately recorded at the same sample rate (16KHz) and resolution (16 bits) on a Sparc 10 computer, and all played through the same playback interface via a pair of headphones. In total, each sentence was listened to at least once in each of the five different formats, and no sentence was heard more than once by each subject.

Test results were analysed by sentence matching software based on dynamic programming. The software attempts to match the sentence perceived by the subject with the known reference sentence and generates the number of correct words, insertions, deletions, substitutions and sums these to give the total number of errors. Although the software does not give an indication of the sources of error, it does provide baseline figures on which we can comment. These figures are shown in table 1 and given in more detail in Appendix A.

	Sentence Format	errors	Intra-listener Std. Deviation
(i)	Original speech Waveform	4.6%	3.5%
(ii)	Synthesised database labels	18.2%	13.3%
(iii)	Synthesised recogniser labels	49.7%	10.7%
(iv)	Synthesised database labels (with pitch)	16.9%	12.9%
(v)	Synthesised recogniser output (with pitch)	43.0%	12.8%

Table 1: Baseline Results of Psycho-acoustic Tests

Figure 1 can be used as a guideline for identifying the points throughout the system where information might be lost or degraded. When listening to the original speech waveforms subjects managed to identify words within sentences at an average rate of about 95%. So even with the original 16-bit 16KHz speech recordings a perception loss (P) of about 5% can be expected.

In sentence format (ii) the hand-labelled phonetic transcriptions of the TIMIT database are converted to a suitable form and then processed by the synthesiser to regenerate a version of the original speech. According to the sentence analysing software, subjects were able to perceive these reconstructed sentences with about an 82% word accuracy. There are four sources of error in generating these sentences; firstly the conversion error (C) from the TIMIT phoneme set to the one used by the synthesiser, secondly the phonetic transcriptions errors (T) that will almost certainly have occurred in labelling the database, thirdly the synthesiser will introduce its own reconstruction errors (S) and lastly the perception loss mentioned earlier.

The tests do show a slight increase in perception ability when pitch is added to the phoneme sequence during re-synthesis as in sentence format (iii), but not significantly since the intra-listener variance is so large. Here subjects were able to perceive sentences with about an 83% accuracy. Including the original pitch contours indeed made the re-synthesised speech more pleasing to the ear, and to a certain extent added some elements of expression and intonation. Unfortunately in these tests the subjects are listening to individual sentences from different speakers and were unable to take advantage of contextual information.

The large vocabulary phoneme recogniser will naturally introduce errors of its own. The phoneme error rates reported for the recogniser used in these experiments are in the region of 30% [Robinson, 1991]. No restrictions have been placed on the database vocabulary or speaker variation, so all results obtained are for a large vocabulary connectionist task with multiple speakers. When the recogniser is brought into the chain the overall word error rate through the system falls to about 50% with no pitch information and to about 43% when the pitch is included. The recognition error shall be referred to as R. Figure 3 shows how the various errors/losses come together in the overall scheme.

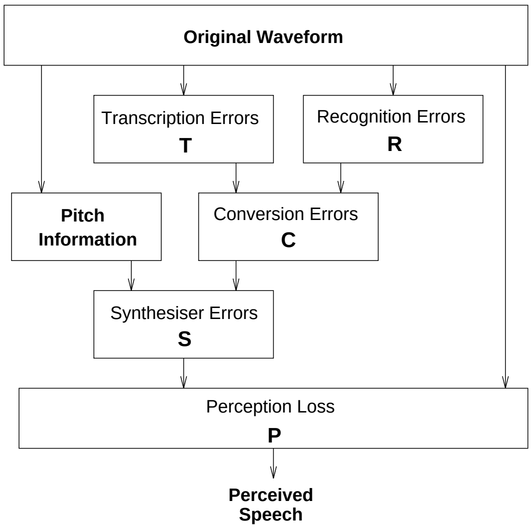

Figure 3: Losses and errors throughout the system

The results of table 1 together with figure 3 suggests that the errors introduced by the synthesiser are not constant but very much dependent upon the amount of errors introduced at earlier stages. The conversion error C, is unknown and difficult to predict but is assumed to be less than 5%.

These initial findings illustrate the complexity of the task whilst giving pointers to the areas that can be improved to give better performance overall. Considering the multiple speaker, large vocabulary, non-contextual task, the results obtained here are very encouraging. With a speaker dependant or limited vocabulary task where the recogniser introduces less errors, the performance can be expected to be much improved. The synthesiser/perception loss appeared to be due to a number of factors, many of which do not greatly affect perception or would be rectified if listening in context. Some of these problem areas are peoples names, specific terminology and function words which were often omitted. A learning curve is also expected in becoming familiar with the synthesiser. The speech rate and changes in voice from sentence to sentence was also found to be a problem.

In the tactile system the synthesiser is replaced by the tactile display and it is expected that the perception loss would be modified according to the mode in which the brain is receiving information.

Intra-listener Variations

In all four of the synthesised sentence formats the intelligibility results from the individual subjects varied widely. Some subjects were able to put together obscure sounding sentences to score highly in the re-synthesised recogniser output whereas others would come up with nothing at all. Some subjects were better at dealing with the longer sentences than others. It is hoped that with sufficient training most subjects would improve in both these areas.

SKIN SENSITIVITY

The sensitivity of the skin is limited to certain frequencies of vibration and lengths of gap detection, which are the limiting factors in passing information via tactile means. This section aims to ascertain whether the skin is able to resolve speech information presented to the skin as spatial points of stimulation representing the individual phonemes.

Hardware Construction

Simple hardware has been constructed for the purpose of investigating the sensitivity of the skin; it consists of two miniature solenoids, which will provide the stimuli, and that have rods which protrude through a flat surface constructed from firm perspex. The two points on the surface are separated by a distance of 12mm and the solenoid rods protrude through the surface a variable distance from 0 to 10mm (set to 1mm in the tests that follow). The rods that make contact with the skin have a cross-sectional area of 1mm^2.

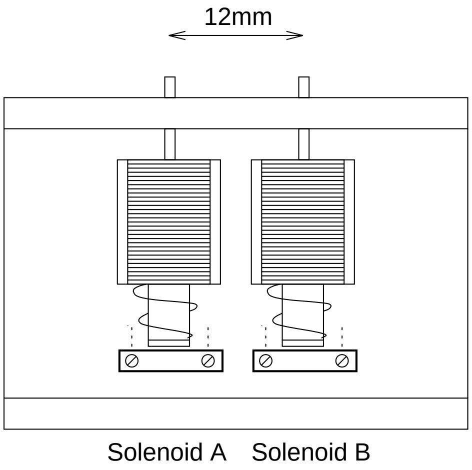

Figure 4: The hardware construction (side view)

It is intended that the forearm is rested on the firm surface to receive a stimulus from each of the solenoids. However, with the hardware construction described above it is just as easy for the finger tips, thenar eminance, palm etc. to be used as the stimulation site instead. The miniature solenoids are driven by digitally synthesised waveforms, configured here to generate pulse and pulse-modulated stimuli.

Waveform parameters such as pulse width, period and modulation rate can be controlled interactively via an on-screen menu, and independent, synchronised waveforms can be sent to each of the solenoids. The waveform generating package described here is also capable of generating analogue signals to drive the miniature solenoids, as is required for certain modes of stimulation.

Mechanoreceptive Systems in the Skin

It is suggested that at least two receptive systems exist in the skin, and that these can be characterised as Pacinian (P) and non-pacinian (NP) types [Sherrick *et al.*, 1990]. The P system is most sensitive to high frequency vibrations, with peak sensitivity in the region of 200-300Hz. The P system also exhibits a rapidly adapting response. The NP system is most sensitive to lower frequencies and is able to localise best at frequencies of 40 - 50Hz, and frequency discrimination is best perceived below 100Hz [Rothenberg*et al.*, 1977]. Thus the frequencies used to stimulate the skin can be carefully chosen to utilise the natural response characteristics of the skin.

There will however be limitations imposed by the hardware on the actual frequencies and coding strategies that can be used to stimulate the skin. The dynamics of any solenoids, transducers etc. that may be used to provide the stimulus will be limited by their mechanical composition. Other restrictions will be imposed to ensure reliable operation and possibly low power consumption.

Despite the unusual frequency response characteristics of the skin most frequencies below about 700Hz [Rothenberg*et al.*, 1977] will give a reasonable sensation without any pain provided that the level of intensity is great enough. But there are clearly a few frequencies of vibration that will provide a clearly defined stimulus over a wide range of intensity levels; this is shown in the experiments of later sections.

The inner ear is an insular organ with an intricate geometric design that accepts a spectral flux of mechanical energy, performs a rapid preliminary analysis of it, then converts it to a nervous message. The skin, by contrast, is a system occupied by myriad entities other than mechanoreceptors and serves many other functions [Sherrick, 1985]. It will be necessary therefore to find the best mode for stimulating the skin, taking into account the other functions performed by the skin and the changing effects that might come about as a result. The skin, although very sensitive, undergoes many cycles of varying acuteness to stimulation making the task of using the sense of touch as a mode of communication that much more difficult.

The other functions performed by the skin often create changes that can reduce or enhance the sensitivity of the skin. For example, when the body becomes overheated blood is pumped around the body and fluids are released externally to regulate the skin temperature. These rapidly changing effects can drastically affect the sensitivity of the skin at a receptor level. Differences will also occur from person to person, with age, time of day, whether its summer or winter etc. [Van Doren, 1990; Verrillo, 1979]. Women are particularly affected by changes in sensitivities during mensuration [Gescheider *et al.*, 1984]. The idea is to use the areas of sensitivity that remain largely unchanged from one set of conditions to another.

Typical frequencies used for tactile displays are in the regions of 30-60Hz and 200-250Hz e.g. [Gescheider, 1990; Rogers, 1970; Grigson, 1989]

Modes of Stimulation

The actual waveforms of vibratory stimuli will affect how well a signal is received at the skin. In its simplest form the waveform will be a sinusoid that is either at full amplitude or off. With digital electronic circuits it is far easier to generate square waves; there is an advantage with square waves in that its fourier composition is of an infinite number of frequencies at multiples of the fundamental albeit with diminishing amplitudes. The square wave frequency can be chosen such that its larger harmonics coincide with the peak sensitivity of the Pacinian receptors in the skin.

The AM signal, and its variants, has been proved a successful method for generating vibratory stimuli [Weisenberger, 1986]. The AM signal, compared with a normal sinewave, exhibits lower energy requirements (necessary for portable applications) and is not a complex signal to generate. This signal also has three discrete frequency components which can be carefully chosen to suitably excite the skin; see figure 5.

Best results are achieved for signal frequencies of around 40Hz with say a 250Hz carrier [Weisenberger, 1986].

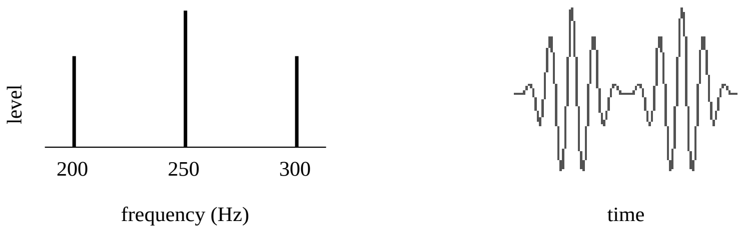

Figure 5: Example - 250Hz sinusoidal carrier amplitude-modulated at 50Hz

Other tried methods of generating suitable stimuli include digitally synthesised pulse envelopes carrying sinewaves or band-limited noise [Van Doren, 1990] and function generated haversine pulses [Sherrick, 1985]. However these are sophisticated signals that are difficult to generate and do not present any significant advantages over the sine, pulse-modulated or amplitude-modulated signals.

Skin Sensitivity Tests

Simple experiments were performed using the hardware set up described above. The reasons for these tests are to determine the absolute data rates of pulse stimuli to the skin that can be reasonably resolved. Searching the whole of the TIMIT database yielded an average phone length of approximately 80ms and only 7% of phones output from the recogniser were of the minimum recognition length of 16ms.

For these simple tests the pulse modulated mode of stimulation was chosen primarily for its simplicity and in practice was found to produce clearly defined sensations on the skin. In all the tests that follow, subjects were able to adjust the intensity of vibrations for maximum comfort but once adjusted was kept constant throughout.

Five subjects took part in these tests which were divided into three distinct sections,

(a) *Testing for frequency definition between 10 and 100Hz:*

Of the two peak sensitivity frequency ranges of the skin, the lower was chosen for these tests. There is much discussion as to which of the two main mechanoreceptive systems in the skin is best able to localise stimulation site. Rogers concluded that frequencies of around 40Hz might provide the best localisation [Rogers, 1970] although it is found that the Pacinian receptors may be responsible for resolving closely spaced temporal events [Van Doren, 1990]. However, it is also found that the high-frequency Pacinian receptors are the one that deteriorate with age [Van Doren, 1990].

In this section of the tests each subject was presented with a single point stimulus on the base of the forearm as a 200ms pulse, as shown in figure 6. The 200ms pulse is an envelope for a frequency burst in the range 10 to 100Hz.

Subjects were able to detect 200ms pulses modulated at frequencies from 10 to 100Hz with no difficulty, although above 80Hz signals were found to be less well defined. All subjects reported the stimulus to be clearly defined and most comfortable between 50 and 70Hz. At lower frequencies it must be considered that the waveform period will start to approach the phoneme data rates that can be expected in the tactile system.

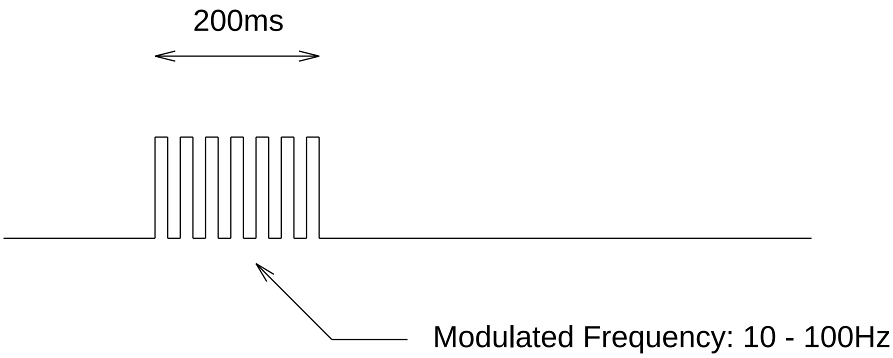

Figure 6: Drive waveform for frequency definition tests

(b) *Minimum pulse width detection:*

Subjects must be able to resolve pulse widths of stimulation that are of the same order as the phoneme rates that can be expected on a tactile display. If the phoneme rate in continuous speech was say ten per second then the average duration of a single point stimulus would be 100ms. Figure 7 shows the waveform type used to generate the stimulus in this section of the tests which was again to just a single point on the skin. The modulated frequency chosen for this section of the tests was 70Hz, as this was one of the better perceived frequencies. Also, the minimum phoneme recognition length is 16ms and a 70Hz signal would produce at least one cycle of oscillation.

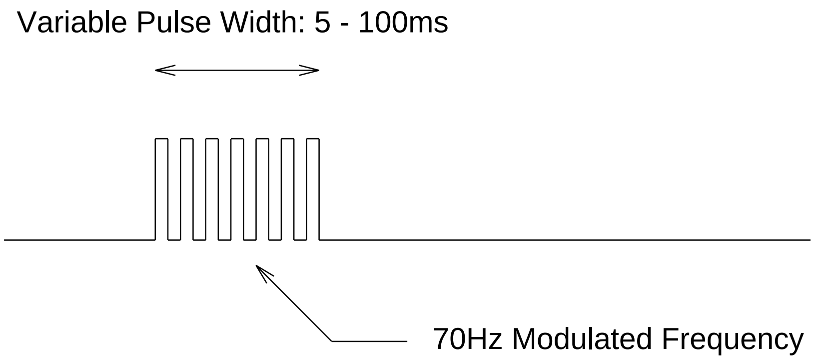

Figure 7: Drive waveform for pulse width tests

Subjects were presented with pulses of random width in the range 5ms to 100ms and were able to detect pulse widths down to about 10ms with regularity. It is noted that when the modulating pulse width falls below the period of the signal being modulated the resulting waveform is merely a single pulse. This did not however seem to affect the subjects ability to perceive the stimulus, but with a high enough frequency there will always be at least one cycle included within the pulse envelope.

(c) Minimum Gap detection between two pulses

On a tactile display it will be necessary to detect sequences of stimuli at different sites on the skin as the phonetic information changes. Between stimuli it may be necessary to have short pauses to assist localisation. The simplest case is to have two stimulation sites on the skin and exciting them successively. The waveforms to achieve this are illustrated in figure 8.

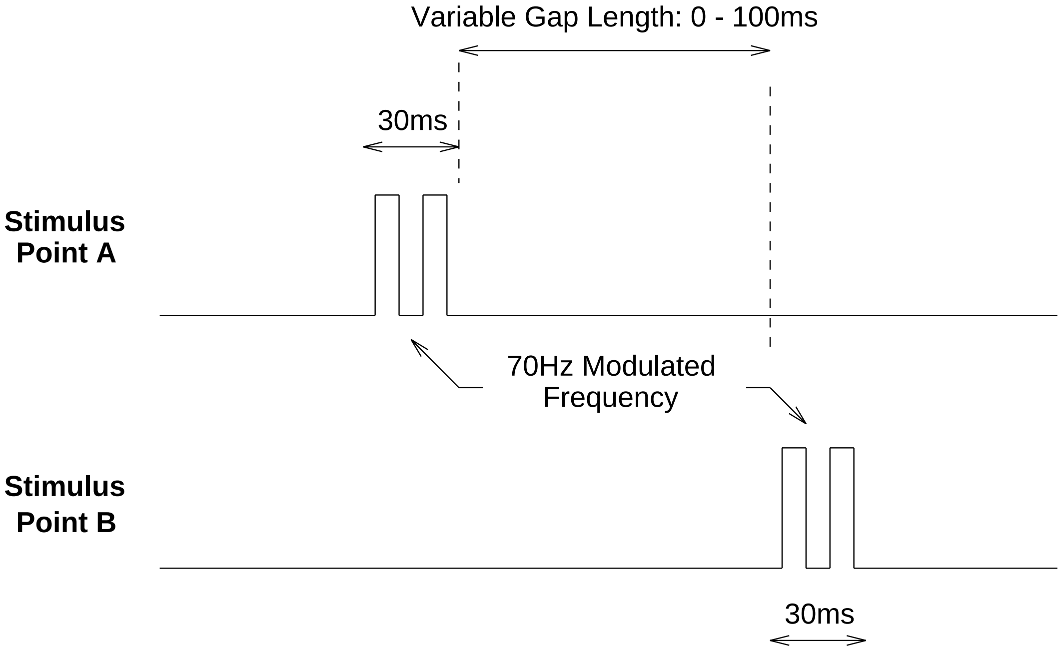

Figure 8: Drive waveforms for gap detection tests

30ms pulses were chosen because this was short enough to easily accommodate the phoneme rates that can be expected and was easily detected by subjects when presented in isolation. The modulated frequency was again 70Hz. The pulse sequence was either as that shown in figure 8 or with solenoid B receiving a signal first on a random basis. The duration between pulses was also varied randomly from 0 to 100ms in 10ms intervals. The subjects were asked to state which site was stimulated first.

With Gap lengths of 10ms and above, all subjects were always able to state which of the pulses was felt first. As the gap between pulses approaches zero it appears that the two stimuli merge and are perceived as a 'phantom stimulus' somewhere between the two stimulation sites. For gap lengths of 10ms and below subjects were able to state the pulse order with an accuracy of about 80% although three of the subjects felt uncertain in making the decision with these short gap lengths.

Putting together the results of the basic tests described above it is possible to have gaps between successive stimuli of about 20ms and excitation durations down to 10ms. If this held true for a full tactile array of say 40 points with a continuous stream of phonetic information then it would be possible to transcribe and present continuous speech information in phonetic form to the tactile display. This area is discussed later in this report.

The general results obtained here are found to agree with the work done by others e.g. [Rothenberg*et al.*, 1977; Sherrick, 1970], and shows that the fundamental rates of pulse signals that can be resolved by the skin are of the same order as phoneme rates in normal speech.

A Full Tactile Grid

What can not be determined from the results of the previous section is how a subject would respond to a full grid of tactile stimulators. It is expected that the full tactile display will consist of 40 to 50 (see following section) stimulators each representing one of the phonemes. There is no obvious way of scaling the results obtained in the tests of the previous section to more than two stimulators. The varying distance between successive stimuli may affect the minimum durations of excitation and with a continuous stream of information the brain will need more time to process and interpret the stimuli.

TWO DIMENSIONAL DISPLAY

The relationship between any two phonemes of a language is a complex one governed by a number of different parameters. To arrive at the required planar format, the multi-dimensional phoneme representation must be reduced to just two-dimensions, whilst maintaining as much of the original parameter information as possible. This can be achieved by performing a proximity analysis and displaying the phonemes according to their similarity. The resulting two-dimensional display will show similar sounding phonemes close together, and dissimilar phonemes spaced far apart. Methods for performing a proximity analysis of this kind are wide ranging and are described at length in [Ellis and Robinson, 1992].

The phonetic tactile system is based on the TIMIT phoneme set which consists of 61 symbols. Table 2 shows the phonemes that are used with the TIMIT database along with a word in which the phoneme appears.

Vowels		Semivowels		Nasals	
eh	bet	l	lat	m	mom
ih	bit	el	bottle	em	bottom
ao	bought	r	ray	n	noon
ae	bat	w	way	en	button
aa	bott	y	yacht	ng	sing
ah	but	hh	hay	eng	washington
uw	boot	hv	ahead	nx	winner
uh	book				
er	bird	Fricatives		Stops	
ux	toot	s	sea	p	pea
ay	bite	sh	she	b	bee
oy	boy	z	zone	t	tea
ey	bait	zh	azure	d	day
iy	beet	th	thin	k	key
aw	bout	dh	then	g	gay
ow	boat	f	fin	dx	muddy
ax	about	v	van	q	bat
axr	butter				
ix	debit	Affricates		Closures	
ax-h	suspect	ch	choke	pcl	pea
		jh	joke	bcl	bee
Others				tcl	tea
h#				dcl	day
pau				kcl	key
epi				gcl	gay

Table 2: Phonemes of the TIMIT Database

The approximate display shown in figure 9 has been generated by self-organising techniques [Ellis and Robinson, 1992], based on a principal component analysis; the display exhibits spatial representation according to phoneme similarities, which is desirable for assisting tactile perception and copes well with small phonetic errors.

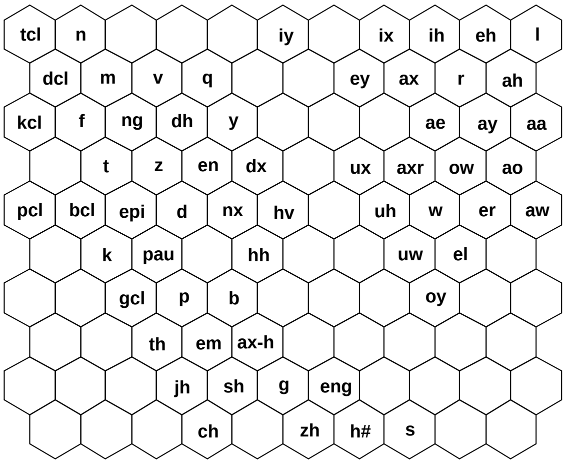

Figure 9: Arrangement of Phonemes on Tactile Display

A Reduced Phoneme Set

It is desirable to reduce the display of figure 9 to a workable size that can be easily implemented in hardware. A reduced phoneme set might also simplify the perception of continuous phonetic information.

The synthesiser used in the earlier tests did not have closures as part of its symbol set. In converting from the symbol set used by the TIMIT database to that used by the synthesiser, it was necessary to merge the closures with their associated stops. With this conversion the synthesiser produced reasonably intelligible speech suggesting that the closures could be a possible area for reducing the phoneme set. Also, the synthesiser only recognises one type of silence (sil) whereas the TIMIT set has three (h#, pau and epi). In the same manner it is possible to reduce three nasals, three vowels and one semi-vowel. It is also noticed that the phonemes considered here for reduction are found to occur seldomly in the TIMIT database. The resulting phoneme set is shown in table 3 and has 46 symbols.

TIMIT	Reduced	TIMIT	Reduced	TIMIT	Reduced	TIMIT	Reduced
h#	sil	ey	ey	z	z	p	p
pau		iy	iy	zh	zh	pcl	
epi		aw	aw	ch	ch	b	b
eh	eh	ax	ax	jh	jh	bcl	
ih	ih	ax-h	ax	th	th	t	t
ao	ao	ow	ow	dh	dh	tcl	
ae	ae	ix	ix	f	f	d	d
aa	aa	l	l	v	v	dcl	
ah	ah	el	el	m	m	dx	
uw	uw	r	r	em	em	k	k
ux		w	w	n	n	kcl	
uh	uh	y	y	en	en	g	g
er	er	hh	hh	ng	ng	gcl	
axr		hv		eng		q	q
ay	ay	s	s	nx			
oy	oy	sh	sh				

Table 3: The reduced phoneme set

With the reduced set of 46 phonemes a smaller grid arrangement of say 7x7 cells can be achieved. With only three spare cells in the grid there is little room for manoeuvre in getting the optimum phoneme arrangement. Instead, an 8x7 grid has been chosen and the phonemes have been arranged to be consistent with the results gained from a proximity analysis of the reduced phoneme set based on principal components [Ellis and Robinson, 1992]. The resulting display is shown in figure 10.

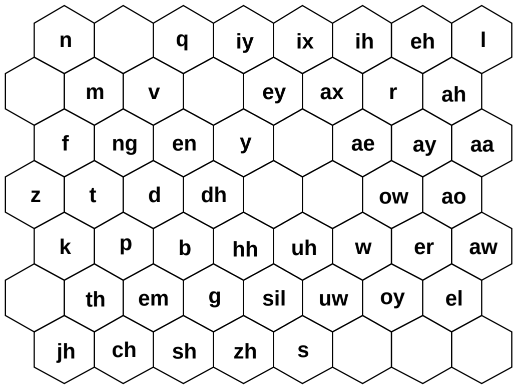

Figure 10: The reduced tactile display

CONCLUSIONS

It has been shown that sentences re-synthesised from a phonetic string can sound reasonably intelligible even when the original pitch contours of the speech are absent. This shows essentially that the phonetic representation maintains a high percentage of the useful information in the original speech. Including pitch information in the re-synthesised sentences made a small difference in the ability to perceive sentences but was found not to be significant. This means that it may not be necessary to include this element of speech information at a tactile level. However, with the pitch information included the resulting speech sounded more natural and pleasing to the ear. It would be desirable to do further studies involving the re-synthesis of recognition results from a database such as the DARPA Wall Street Journal based Continuous-Speech Corpus [Paul and Baker, 1992]. The contextual nature of the database will be similar to the information that will be processed by a tactile aid and thus be a more accurate assessment of the tactile performance that can be expected.

Studies and experiments on the sensitivity of the skin show that the data rates associated with the phonetic format are similar to the rates that can be processed through the sense of touch. The results obtained here give a rough idea of how a full tactile display will perform but a multi-point stimulator will have to be constructed for an accurate assessment of the full tactile display.

If the brain is able to process continuous tactile information at the same rates at which phonemes appear in natural speech, then the sections described in this report could form the basis of a complete tactile hearing system.

ACKNOWLEDGEMENTS

The base phone probabilities of the work described in this report has been supplied from work carried out under the ESPRIT WERNICKE project (6487). The authors would also like to thank members of the Speech, Vision, Robotics and AI group of Cambridge University Engineering Department especially Christine Tuerk who supplied the pitch

tracking information used in some of the experiments and Richard Prager for his useful knowledge of waveform generating packages.

Appendix A
Detailed Results of Psycho-acoustic tests

Sentence Format		errors by each listener			
		1	2	3	4
(i)	Original speech Waveform	0.0%	2.0%	7.8%	7.4%
(ii)	Synthesised database labels	18.9%	15.7%	7.0%	3.6%
(iii)	Synthesised recogniser labels	72.2%	36.7%	44.6%	45.3%
(iv)	Synthesised database labels (with pitch)	29.1%	10.5%	2.0%	0.0%
(v)	Synthesised recogniser output (with pitch)	43.6%	35.7%	43.1%	41.8%

Table 4: Results for subjects 1 to 4

Sentence Format		errors by each listener			
		5	6	7	8
(i)	Original speech Waveform	1.8%	2.0%	9.1%	1.9%
(ii)	Synthesised database labels	12.2%	16.1%	24.1%	34.5%
(iii)	Synthesised recogniser labels	58.0%	47.1%	54.2%	50.9%
(iv)	Synthesised database labels (with pitch)	17.9%	42.0%	16.4%	24.1%
(v)	Synthesised recogniser output (with pitch)	35.3%	24.6%	71.7%	55.9%

Table 5: Results for subjects 5 to 8

Sentence Format		errors by each listener			
		9	10	11	12
(i)	Original speech Waveform	0.0%	0.0%	0.0%	1.2%
(ii)	Synthesised database labels	6.9%	46.0%	8.2%	7.4%
(iii)	Synthesised recogniser labels	56.4%	36.8%	42.3%	40.4%
(iv)	Synthesised database labels (with pitch)	7.5%	21.6%	3.7%	6.5%
(v)	Synthesised recogniser output (with pitch)	38.9%	38.8%	38.1%	37.8%

Table 6: Results for subjects 9 to 12

REFERENCES

[Weisenberger and Russell, 1989] J.M. Weisenberger and A.F. Russell (1989), *Comparison of two single-channel tactile aids for the hearing-impaired*, Journal of Speech and Hearing Research 32, 83-92.

[Weisenberger and Broadstone, 1989] J.M. Weisenberger and S.M. Broadstone (1989), *Evaluation of two multichannel tactile aids for the hearing impaired*, J. Acoust. Soc. Am. 86(5), 1764-1775.

[Brookes and Frost, 1983] P.L. Brookes and B.J. Frost (1983), *Evaluation of a tactile vocoder for word recognition*, J. Acoust. Soc. Am. 86, 1764-1775.

[Sherrick, 1985] C.E. Sherrick (1895), *Touch as a communicative sense*, J. Acoust. Soc. Am. 77(1), 218-219.

[Robinson, 1992] A.J. Robinson (1992), *A real-time recurrent error propagation network word recognition system*, In Proc. ICASSP-92(3), pages 617-620.

[Lamel *et al.*, 1987] L.F. Lamel, R.H. Kasel and S. Seneff (1987), *Speech Database Development: Design and Analysis of the Acoustic-phonetic Corpus*, In Proceedings of the DARPA Speech Recognition Workshop, pages 26-32.

[Ellis and Robinson, 1992] E.M. Ellis and A.J. Robinson (1992), *Two dimensional representation of phonemes of the English language*, In Proc. IOA Conference on Speech and Hearing 14(2), pages 407-414.

[Allen *et al.*, 1987] J. Allen, M. Hunnicutt and D. Klatt (1987), *From text to speech: The MITalk system*, Cambridge University Press.

[Tuerk, 1993] C. M. Tuerk (1993), *Automatic speech synthesis using auditory transforms and artificial neural networks*, PhD Thesis, Cambridge University Engineering Department, 1993.

[Robinson, 1991] Tony Robinson (1991), *Several Improvements to a Recurrent Error Propagation Network Phone recognition System*, Technical Report CUED/F-INFENG/TR.82, Cambridge University Engineering Department, 1991.

[Sherrick *et al.*, 1990] C.E. Sherrick, R.W. Cholewiak and A.A. Collins (1990), *The localisation of low and high frequency vibrotactile stimuli*, J. Acoust. Soc. Am. 88(1), 169-179.

[Van Doren, 1990] C. L. Van Doren (1990), *Vibrotactile temporal gap detection as a function of age*, J. Acoust. Soc. Am. 87(5), 2201-2206.

[Verrillo, 1979] R. T. Verrillo (1979), *Change in vibrotactile thresholds as a function of age*, Sens. Process 3, 49-59.

[Gescheider *et al.*, 1984] G. A. Gescheider, R. T. Verrillo, J. T. McCann, E. M. Aldrich (1984), *Effects of the menstrual cycle on vibrotactile sensitivity*, Perception and Psychophysics 36(6), 586-592.

[Gescheider, 1990] G. A. Gescheider (1990), *Vibrotactile intensity discrimination measured by three methods*, J. Acoust. Soc. Am. 87(1), 330-338.

[Rogers, 1970] C. H. Rogers (1970), *Choice of stimulator frequency for tactile arrays*, IEEE Trans. on Man-machine Systems MMS-11(1), 5-11.

[Grigson, 1989] P. J. W. Grigson, R. A. Giblin(1989), *The hand-tapper: A communication and information system for the deaf-blind using the British Manual Alphabet*, RESNA 12th Annual Conference, New Orleans, Louisiana, p139-140.

[Weisenberger, 1986] J. M. Weisenberger (1986), *Sensitivity to amplitude modulated vibrotactile signals*, J. Acoust. Soc. Am. 80(6), 1707-1715.

[Sherrick, 1985] C.E. Sherrick (1985), *A scale for rate of tactual vibration*, J. Acoust. Soc. Am. 78(1), 78-83.

[Rothenberg*et al.*, 1977] M. Rothenberg, R.T. Verrillo, S.A. Zahorian, M.L. Brachman and S.J. Bolanowski Jr. (1977), *Vibrotactile frequency for encoding a speech parameter*, J. Acoust. Soc. Am. 62, 1003-1012.

[Sherrick, 1970] C.E. Sherrick (1970), *Temporal ordering of events in haptic space*, IEEE Transactions on Man-machine systems, Vol. MMS-11(1), 25-28.

[Paul and Baker, 1992] D. Paul and J. Baker (1992), *The Design for the Wall Street Journal-based CSR Corpus*, DARPA Speech & Nat. Lang. Workshop, February 1992.