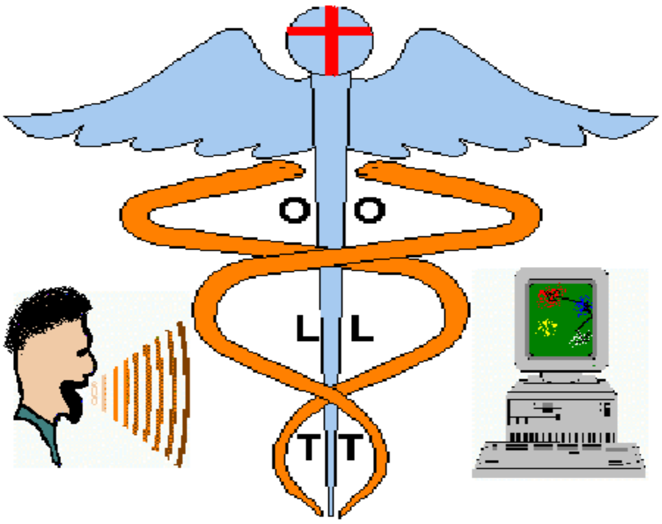
Optical Logo-Therapy (OLT)

Optical Logo-Therapy (OLT)

Computer-Based Audio-Visual Feedback Using Interactive Visual Displays for Speech Training



Athanassios Hatzis

October 1999

Submitted in accordance with the requirements for the degree of Doctor of Philosophy to Department of Computer Science, University of Sheffield.

Abstract

In this thesis we present Optical Logo-Therapy (OLT), a real-time computer-based audio-visual feedback for speech training, in which the speech acoustics are transformed into visual events on a two-dimensional display called a 'phonetic map'. *Phonetic maps* are created by training a neural network to associate acoustic input vectors with points in a two-dimensional space. The target points on the map can be chosen by the teacher to reflect the relationship between articulatory gestures and acoustics. OLT thus helps the client to become aware of, and to correct, errors in articulation. *Phonetic maps* can be tailored to different training problems, and to the needs of individual clients. We formulate the theoretical principles and practical requirements for OLT and we describe the design and implementation of a software application, OLTK. We report on the successful application of OLT in two areas: speech therapy and second language learning.

Acknowledgements

This thesis is the product of four years personal work and as every other thesis it might never have come to an end without the help of many people I happened to meet and know during this period. Moreover, it is essential for one to manage and maintain a good balance in every aspect of his/her life; as such, I would like to present you all those that contributed the most in those four years of my life, and at the same time express my feelings for their help.

First I would like to express my gratitude to my supervisors **Phil** and **Sara**, for their guidance, encouragement, scientific integrity, and organisation of the experiments. I also wish to express my warmest thanks to: **Kate** for embracing with enthusiasm the application of OLT software, the parents and their children participated in the treatment program with OLT, and the junior schools of Sheffield that helped me with the recordings of their students. In addition, I am grateful to **Barnsley College** for the studentship I took on the first year of my studies.

Many thanks to my colleagues: **Miguel** for his kindness to do the tutorials on statistics, **Vinny** for his precious offer of a neural network piece of code and for helping me to develop my own, **Gethin** for sharing some of his knowledge on speech recognition, **Andy** and **Yoshi** for their patience to explain some of the maths involved in my research, **Richard** for his advice and demonstration of phonetics with OLT, and last our support team for helping with compilation, debugging and installation of programs. I also feel particularly fortunate for the nice and friendly atmosphere created by the members of our group, and especially for the good company of **Brian** and **Vinny**.

Furthermore I owe a dept of gratitude to all the other friends of mine I made in U.K for the joyful moment we had together: **Ahmet, Pantoula, Natia, Aggeliki, Michalis, Christiana, Michalis, Stella, Georgina, Harris, Lefteris, Markos, Emilios, Fotis, Petros, Petros, Meletis, John, Evi, David, George, Thanassis, and Fotini**. Special thanks to my close friends **Zissis, Kyriakos, Konstaninos, Stefanos,** for their good company; and many thanks also to my good friends **Stefanos, Dimitris, Stavros**, **and my father John,** for participating in my experiments with OLT.

Lastly, and most particularly, I want to thank **Fr. John** for his spiritual guidance and for the proof reading of my thesis. Also I feel much obliged to the Greek community of Nottingham (**Fr. Iakovos, John and Elisabeth, Panagiota and Dimitris, Georgia, Iakovos, and Antreas, Dimitra, Ector and Antreas, Helen and Kyriakos, Helen and Marios**) and the Greek community of Edinburgh (**Fr. John, Ian and Noula, Sofia**) for their help all these years.

Dedication

I kept separate the heroes behind the scene to refer to. These are my parents; **John Hatzis** and **Christina Hatzis,** and this thesis is dedicated to them. They have struggled all these years in order to contribute substantially to my financial burden. Without a doubt, these are the persons they love me more than anyone else in this world and that is why I love them too. This dedication is currently the least I can do to please them, but I give them my word that I will try to care about them as much as I can in the rest of their lives.

Finally it would be a vast omission not to mention my refugee, my consolation, my protection, the **mother of God** and mother of us all and ever virgin **Mary**, our **Panagia**. She brought me back to her Son, **Christ my God**, my Creator, my Saviour, my Light, my Inspiration, without whom I would have been lost. All the glory, all the honour are for Him and only for Him, *"for every benefit and every perfect gift is from above and cometh down from Thee, the Father of Lights; and to Thee we ascribe glory and thanks and worship, to the Father and to the Son and to the Holy Spirit, now and for ever and from all ages to all ages (liturgy of St. John Chrysostom, dismissal)"*. Amen.

Declaration

This dissertation is the result of my own work and is not the outcome of any work done in collaboration. I hereby state that this dissertation is not substantially the same as any I have submitted for a degree, diploma or other qualification at any other University. Further, no part of this dissertation has already been or being concurrently submitted for any such degree, diploma or other qualification.

Abbreviations

ADC	= Analog to Digital Converter
AM	= Accent Modification
ANN	= Artificial Neural Network
DFT	= Discrete Fourier Transform
DSP	= Digital Signal Processing
EMA	= Electromagnetic Articulography
EPG	= Electropalatography
EVF	= Effective Visual Feedback
EZWGL	= Easy Widget Graphics Library
FAD	= Functional Articulation Disorders
FIFO	= First In First Out
HARP	= Speech Rehabilitation Aid for the Hearing Impaired
HDF	= Hyperquadratic Discriminant Function
HMM	= Hidden Markov Model
HTK	= Hidden Markov Toolkit software
I/O	= Input/Output
IDFT	= Inverse Discrete Fourier Transform
ISTRA	= Indiana Speech Training Aid
KNN	= K-Nearest Neighbour
KP	= Knowledge of Performance
KR	= Knowledge of Results
LEDA	= Library of Efficient Data types and Algorithms
LPC	= Linear Prediction Analysis
LVQ	= Linear Vector Quantisation
MAAP	= Motoric Atomisation of Articulatory Performance
MAP	= Maximum a Posteriori
MRI	= Magnetic Resonance Imaging
MSECT	= Mean Square Error Coordinate Transformation

MSSE	= Mean Sum-Squared Error
NCVS	= National Centre for Voice and Speech
NLLT	= Non-linear Linear Transformation
OLT	= Optical Logo-Therapy
OLTK	= Optical Logo-Therapy Toolkit
OOTS	= Out Of Training Space
PCA	= Principal Component Analysis
PDF	= Probability Density Function
PPANN	= Posterior Probabilities Artificial Neural Network
SAMANN	= Sammon Artificial Neural Network
SFS	= Speech Filing System
SILDETECT	= Silence Detection software
SMANN	= Sigmoid Mean sum-squared Artificial Neural Network
SNNS	= Stuttgart Neural Network Simulator software
SOM	= Self-Organising Maps
SPEAKALOOK	= Speak And Look
SPEDATO	= Speech and Data Tool software
STDIN	= Standard Input
STDOUT	= Standard Output
VATA	= Vowel Articulation Training Aid
VSA	= Visual Speech Apparatus

Chapter 1

In the beginning was the ***Logos****, and the* ***Logos*** *was with God, and the* ***Logos*** *was God. He was in the beginning with God; all things were made through Him, and without Him was not anything made that was made.* ***St. John-Gospel (1:1-3) (Translation-RSV)***

By faith we understand that the world was created by the ***Logos*** *of God, so that what is seen was made out of things which do not appear.* ***St. Paul-Hebrews (11:3) (Translation-RSV)***

We also, by God's grace, briefly indicated that the ***Logos*** *of the Father is Himself divine, that all things that are owe their being to His will and power, and that it is through Him that the Father gives order to creation, by Him that all things are moved, and through Him that they receive their being (1:1). He made all things out of nothing through His own* ***Logos****, our Lord Jesus Christ (1:3).*

St. Athanasius-On the incarnation (Translation-C.S.Lewis)

Introduction

This thesis is concerned with computer-based speech training. In recent years there has been considerable interest in the possibility of technological help in this important and human resource-hungry field. Though a number of software applications for speech training have appeared, their functionality still leaves much to be desired. We begin by reviewing how speech training experts work and what problems they face.

1.1 The feedback problem in speech training

Two groups of people that require speech training are: foreign language learners, and individuals with speech-language disorders. What these groups share is their deviance in production when compared with the ambient language (the norm). Let us take two typical examples one for each group.

Suppose you know a foreigner that wants to learn English as a second language. It is usually the case that unless this speaker has some practice in listening and speaking the language from/with other English native speakers s/he will not be able to obtain the accent of the language. It is not also certain that even after continuously practising English pronunciation the student will manage to master and maintain the functional skills required for perfection. You can imagine a similar scenario, this time with an adult person having a lisping problem, pronouncing /s/ as /θ/. Under normal conditions it is highly unlikely that this person will be able to correct his/her mispronounced words simply by self-monitoring. Assuming that there are no mental or physical reasons it is worth while to ask why the speakers cannot reach the norm in speech production. Typical issues that reveal the tip of the iceberg can be :

(a) How the speaker becomes aware of the articulatory or perceptual characteristics of the production which s/he is aiming at.

(b) How the speaker recognises the deviancies in speech production.

(c) How the speaker judges the accuracy of the pronunciation which s/he is aiming at.

(d) How the speaker learns control of the speech mechanisms in order to reach productions that objectively can be judged as good, or correct.

(e) How s/he manages to maintain learned skills and obtain stable speech production.

It is the specific treatment method applied in speech training (either for therapy or for improving competence, as in foreign language pronunciation) that attempts to give answers to these questions. Perhaps the most important factor (and the most obvious one at a quick examination of the problem) leads us to think about the inadequate self-monitoring abilities of the speaker for appropriate speech production. This is particularly evident with hearing-impaired populations : Mahshie, **[63]**, reports that *"without adequate "built-in" feedback, the individual must increasingly rely on alternative sources of speech information"*.

1.2 The speech processing model

The last argument raises two important questions: first what kind of internal information the client has access to, and second what kind of alternative information we can provide him with. Although we acknowledge that this topic is on its own extremely difficult in nature and goes beyond the scope of this thesis, we recognise the importance of linking the feedback method of our application with a well established theoretical model of human speech processing.

1.2.1 <u>The 'speak-hear' learning cycle</u>

Apparently the predominant sensory channel for the development and correction of speech is hearing. It is a fact that the speak-hear learning cycle (**Figure 1-1**) of a human is a natural process and becomes fully developed from the very early stages of our life.

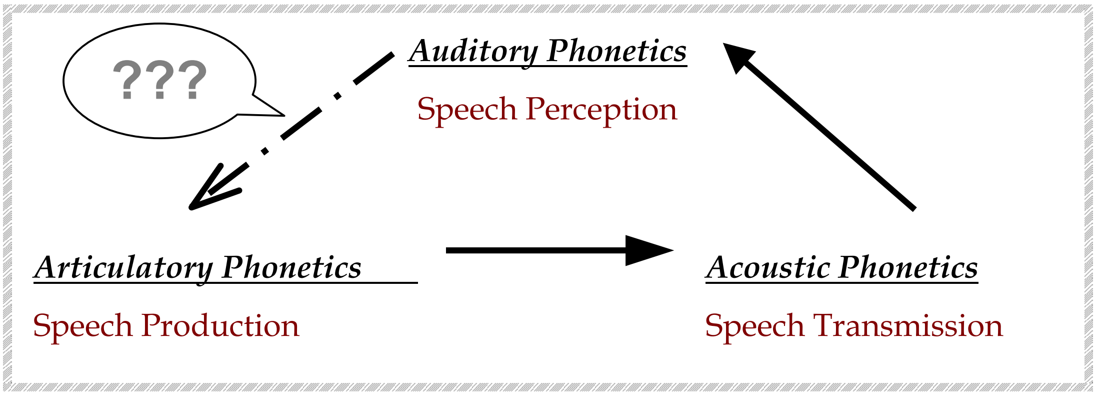


Figure 1-1 *Phonetics science and the speak-hear learning cycle*

As Mahshie states, **[63]**, *"Through feedback of their production attempts, children learn to adjust and modify their articulatory patterns until the utterance serves the intended function. Thus feedback is the mediator between desired outcomes and production"*. Learning to speak happens on a trial-and-error basis through attempting to imitate the perceived sounds. At the same time a person tries to establish cognitively and unselfconsciously an association between the acoustics and the mechanisms of speech production together with other external (environment) and internal (knowledge base) sources of information related to the meaning of the perceived utterance.

Although much research has been devoted to the field of speech perception the last few years, understanding of this complex process is still limited. It has been argued, **[61]**, **[60]**, that the processes of speech perception and production are innately linked (**Figure 1-1**). It is claimed that this link between perception and production, know as the motor theory of speech perception, can account for some of the most complex processes described in the perceptual literature, such as cue trading, duplex perception and findings from studies of audio-visual speech perception. However, these claims are not currently backed by detailed experimental evidence and remain little more than interesting hypotheses.

1.2.2 The psycholinguistic perspective

A more recent approach that takes into account many more factors, such as cognitive skills, for the normal development of speech in children, and has influenced considerably the teaching of speech skills and therapy is the psycholinguistic perspective, **[77]**. The *"essence of the psycholinguistic approach is the assumption that the child receives information of different kinds (e.g. auditory, visual) about an utterance, remembers it and stores it in a variety of lexical representations (a means for keeping information about words) within the lexicon (a store of words), then selects and produces spoken and written words"*, **[96]**, (**Figure 1-2**).

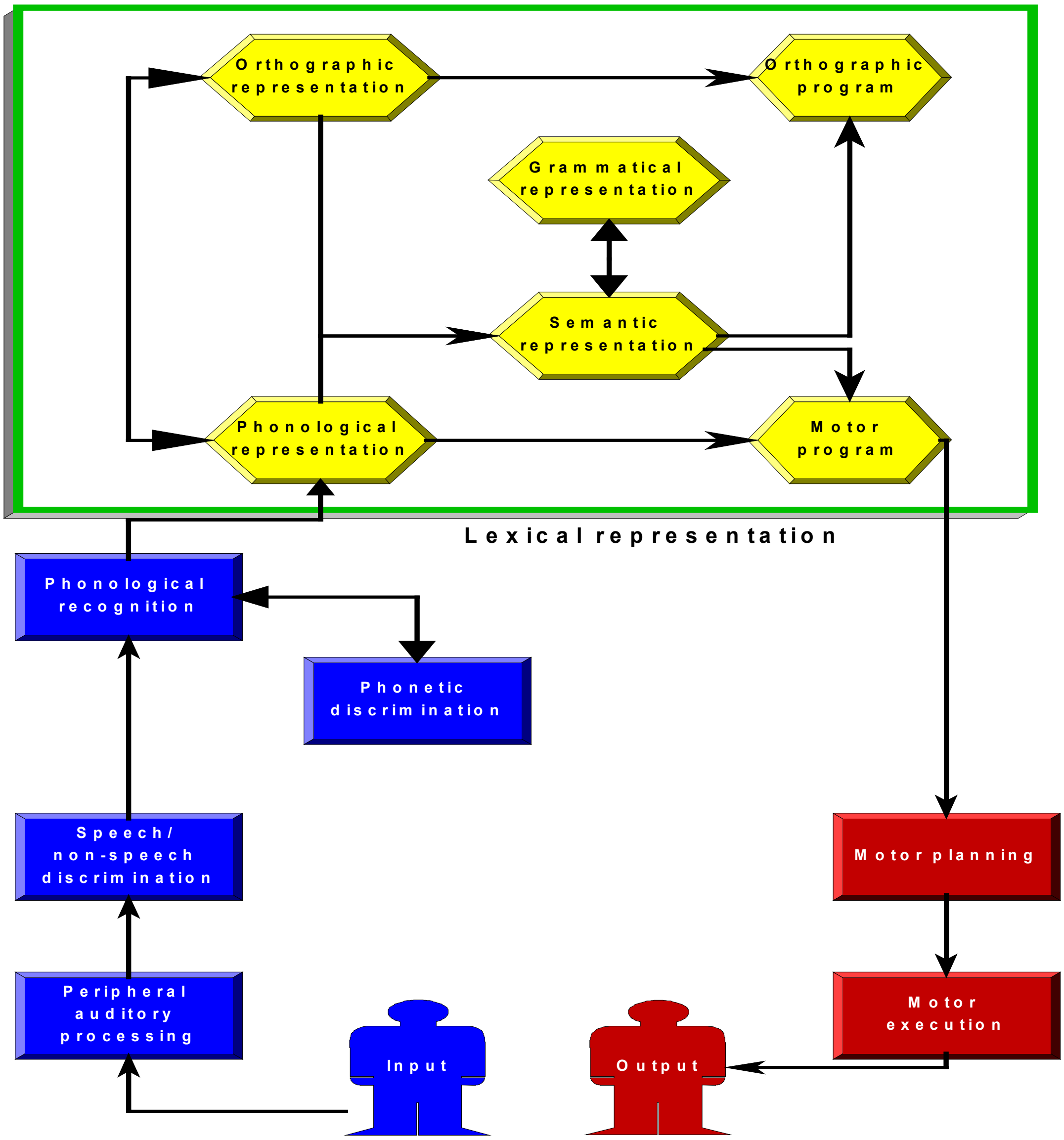


Figure 1-2 *The psycholinguistic model: blue labels indicate the input mechanisms, red labels the output mechanisms, and the yellow labels indicate the lexical representation. (Copied and modified from [96]).*

1.3 Speech training according to etiologies and symptoms

At this point one should ask where in the psycholinguistic model speech processing may break down. This is a very important question, because effective speech training is based first of all on the understanding of the etiological factors and symptoms of the case. Speech defects, according to the model we described, may occur closer either in the input, lexical representation or the output levels of the speech processing. Before we continue the analysis of the type of speech defects in relation to the psycholinguistic model we introduce two other perspectives of classifying speech problems, **[96]**, that will help us analyse the cases we examine in this thesis.

1.3.1 The medical perspective

According to the medical perspective, and if we take also into account the etiological distinction, we can classify speech disorders as organic and functional, depending on whether or not a given disorder is caused by identifiable abnormalities of an organ of the body, **[12]**. In this thesis we deal with a specific type of speech disorder known as 'functional articulation disorders', **[35]**, (FAD) and in particular we examine the case of lisping. Although accent modification (AM), does not constitute a speech disorder, the type of segmental errors in speech production are quite often similar with those of *FAD* and certainly we assume a normal functioning of the organs of the body related to speech.

1.3.2 The linguistic perspective

It is the linguistic approach that broadens the way practitioners conceptualise speech problems. This approach takes into account the symptoms that appear. According to Stackhouse and Wells, terms like *phonetic* or *articulatory* difficulties emphasise that the person appears to have difficulties with the production of sounds,

[96]. Same authors use the term *phonological* as the *"ability to use speech sounds appropriately to convey meaning. This failure to signal contrast between words is known as a phonological problem."* Without being very strict in terminology, we can describe the lisping problem as phonetic while the *AM* case is both phonetic and phonological. Phonetic because of the problem of how to produce correctly the sounds of the foreign language and phonological since there is lack of phonological awareness of where and how these sounds are used in the foreign language. Another distinction we would like to make clear is between prosodic and segmental problems. These can appear in both of the previous types to refer to features of speech such as rhythm, stress, intonation, as opposed to the levels of segmental analysis, such as words, syllables, phonemes, phones, or phonetic features. In this thesis we are focusing on segmental phonetic characteristics and voicing.

1.3.3 The psycholinguistic perspective

The psycholinguistic approach attempts to consider in more detail the underlying cognitive processes. In particular it seeks to find the problematic links between the mental or lexical representation and the speech perception, and speech production levels.

I Psycholinguistic explanation for the lisping problem

The lisping problem we described as a functional articulation disorder breaks down in the motor program level of the lexical representation. According to Stackhouse and Wells, **[96],** *"the motor program for a word consists of a series of gestural targets for the articulators (tongue, lips, soft palate, vocal folds), stored in the lexical representation, that are designed to achieve an acceptable pronunciation of the word"*. It is the lingual gesture that appears to be problematic in the case of lisping. On the contrary the motor execution level according to the same authors concerns anatomical problems with any part of the vocal track which in our case as we have stated in **§1.3.1** are absent.

II Psycholinguistic explanation for the AM case

The psycholinguistic explanation for the *AM* case appears to be rather more complicated because the speech processing breaks down in more than one level of the model. First we have to consider the phonetic discrimination level. Stackhouse and Wells, **[96],** state that *"when the input contains novel phonetic material, e.g. from an unfamiliar accent of English, or from an unfamiliar language, the child needs to be able to sort out the unfamiliar material in phonetic terms, for example by mapping the percept onto new articulatory routines in an attempt to reproduce the unfamiliar sound"*. Following this reasoning we can have a partial explanation for the mapping of the English sibilant fricatives /s/ and /ʃ/ into the single Greek /s/ sound caused by a Greek native speaker learning English. Once this wrong mapping has been established, inside the lexical representation, we have a wrong phonological representation. The reason for that is that usually there is no stored sound contrast between /s/ and /ʃ/ for a Greek speaker because such phonemic contrast does not appear in the orthographic or phonological system of the Greek language. Consequently, a wrongly stored phonological representation results in a wrong motor program therefore the attempted production of these sounds is wrong. It worth noticing here that the intact orthographic, grammatical and semantic representation allows understanding of speech and communication to continue effectively despite the other problems. It is perhaps for these reasons that pronunciation errors are not given much attention by the speaker of a foreign language. Nevertheless, there are cases that these extra mental queues are not sufficient for speech understanding and communication, consider for example a Greek speaker trying to understand the utterance *"she sells, sea-shells"* spoken fast by an English speaker in isolation and consider also the attempt to deliver this spoken utterance to some other English speaker.

1.4 Augmented self-monitoring feedback in speech training

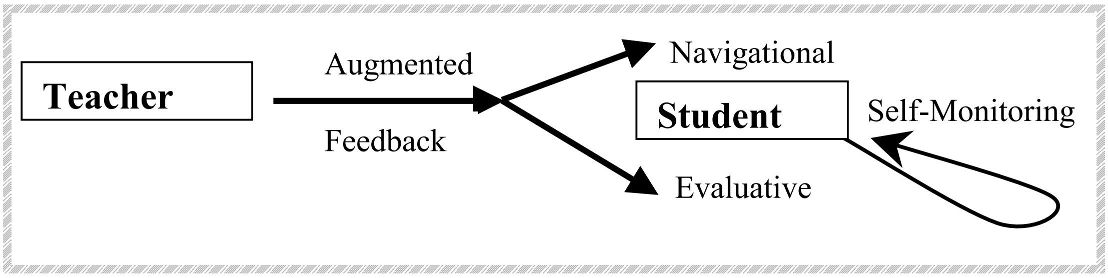


Figure 1-3 *The feedback relationship between the Teacher and Student*

Speech training is an area in which the teacher should try hard to understand the relationship that exists between the audible errors and the inaccurate control or position of various articulators in order to deliver augmented feedback to the self-monitoring abilities of the student (**Figure 1-3**).

Non-instrumental treatment methods have been based on other than hearing forms of augmented feedback, like looking at certain articulatory mechanisms such as lips, jaw or tongue (visual feedback), or feeling the therapist's or their own face, throat and expiration air (tactile), **[73]**. Whatever the form of this feedback, the aim is to make the client aware of the characteristics of target production, and navigate him/her with aural instructions on how to reach and maintain correct speech production, thus navigational feedback.

Moreover it had always been taken for granted that the improvement of the client's speech involves in addition an evaluative form of feedback that the therapist tries to deliver together with a sort of reinforcement in a series of attempts. Therefore the self-monitoring abilities of the learner are enhanced through two types of augmented feedback; navigational and evaluative one (**Figure 1-3**). Ruscello refers to these terms as knowledge of results (KR) and knowledge of performance (KP) and speaks about instrumental treatment. *"The immediate feedback of the performance signal via KR and KP provides an opportunity for cognitive analysis of the desired behaviour. This information is used to verify response accuracy and as necessary, furnish cues for the correct production of succeeding trials."*, **[92]**.

1.5 Non-instrumental articulation remediation-treatment models

Let us recapitulate the problem so far. We want to apply a certain speech treatment methodology that provides augmented feedback to a client appearing to have segmental phonetic problems in *FAD* or *AM* cases. This kind of feedback should try to make the client realise and fix the problematic association that exists between the phonological representation and the motor program. Several treatment programmes, very popular at least till the early 1980s, embody important principles that are used even today in modern speech training methods. These fall into three major categories according to their theoretical bases, and are : behavioural learning theory, theory of skilled movement acquisition and the phonological theory of distinctive feature analysis. We make the assumption here that the teacher is not using any technological aids but that treatment is solely made on a one-to-one basis. We will call these methods non-instrumental articulation remediation-treatment models.

1.5.1 Skilled movement acquisition-McDonald's sensory-motor approach

This treatment is based upon the theory that articulation is basically a motoric activity that is served by sensory feedback, [67]. The main idea here is that during production the learner is instructed to feel for the articulatory contacts (tactile feedback), to report the direction of articulatory movement (proprioceptive feedback), and to listen to the auditory aspects of the target sounds. In general the learner reports the sensations associated with these articulations. The treatment stages appear to be similar to those of Van Riper's method.

1.5.2 Phonological function - Blache's Distinctive feature analysis

A distinctive feature approach organises sounds into classes based on shared acoustic and articulatory features; the hoped-for result of treatment is that generalisation will occur from the treated sounds to other sounds that share similar features and inside utterances of increasing complexity. The basis phonological function is contrast. In Blache's approach, **[11]**, sounds are taught as contrasting units within words using contrast therapy activities. According to Pamela Grunwell, **[32]**, it is the features, not the phonemes, that are the ultimate or minimal distinctive sound units in phonological analysis and in the sound patterns of languages. The idea of separating out features of sounds is very important for the description of a number of phonological processes. The latest theories of non-linear phonology based on syllables still make wide use of them, **[95]**.

1.5.3 Behavioral learning - Van Riper's therapy model

The best known, and once very popular, remediation-treatment model in speech training is Van Riper's behavioural learning, [98]. According to Hoffman and Schuckers, [45], "*this approach to treatment is based upon a particular form of behavioural learning known as instrumental reward conditioning in which the teacher presents a set of stimuli intended to evoke the desired response. If the learner's response is judged to be adequate, the teacher presents a reward. The reward is a stimulus set judged to be pleasant and intended to increase the strength of the relationship between stimuli and response".* The treatment sequence includes three main stages :

(a) Awareness of the characteristics of desired speech production

(b) Awareness of the differences between the desired production and the erroneous production.

(c) Correct production of utterances of increasing complexity (isolated sounds, syllables, words, sentences, conversational speech).

1.6 Our theoretical framework on treatment methodology

More modern approaches such as the motoric atomisation of articulatory performance (MAAP), [45], show that a treatment methodology is required to cover the interrelationship between various articulation treatment theories, and to combine certain important aspects of them. According to Ruscello, **[92]**, treatment methods can be broadly categorised into two groups, Motor and Cognitive-Linguistic. The first is designed *"to teach the complex motor skill movement sequences required in the production of a sound"*, the second *"targets phonological rules which are thought to be missing from the child's phonological rule system"*. The latter *"incorporates phonetic* *production in practice units designed to facilitate the acquisition of phonological rules. For example, many Cognitive-Linguistic approaches use minimal contrast techniques to present phonemic contrasts."*

In this thesis we examine only the feedback element of an instrumental treatment method. However the theoretical framework we adopt to be the basis for building our computer-based system embodies a combination of important treatment principles borrowed from the aforementioned non-instrumental approaches:

1.6.1 Visuomotor tracking

In teaching correct articulation of speech sounds to the client the teacher has to *"give instructions on how to produce them, describe tongue positions, lip configurations, direction of airflow, mouth opening or closing, articulatory movements and contacts needed to produce the target sounds… such information given back to a person about his or her neurophysiological activities through a mechanical device is called biofeedback"*, **[41]**. This type of feedback is what we need to provide our client to assist his/her audible awareness and the correction of the relationship between the acoustics and the segmental articulatory characteristics of his/her speech production. The student has to become aware of and report if necessary directions of the movement of articulatory organs such as lips, jaw, and tongue. Watson and Kewley-Port, **[99]**, state that *"the ultimate goal of speech training is that the trainee internalise the information provided in feedback. He or she must eventually learn to attend to the proprioceptive cues associated with the production"*. Effectiveness here occurs when that feedback is related instantaneously with the acoustics of the speech production coupling the auditory feedback through hearing perception. Davis and Drichta, **[17]**, report that *"biofeedback derives its effectiveness by making ambiguous internal cues explicit, thereby providing accurate information about changes in target responses during training so that instrumental control of the response is facilitated"*.

A special case of biofeedback satisfying the characteristics of the feedback we described is best met at the 'visuomotor tracking', **[107]**. *Visuomotor tracking* tasks are based on the principle that some physical measure reflecting movement performance is fed back visually in real-time. Ziegler et al, states that, *"as tracking tasks contain, as an* *essential component, a biofeedback condition, and as biofeedback is supposed to enhance the learning of novel motor skills, visuomotor tracking has become an important paradigm in motor learning research"*, **[107]**.

1.6.2 Visual contrast

We can distinguish between patterns of normal sound production and patterns of misarticulations. According to Hegde and Davis, **[41]**, selection of patterns can be based on place-manner-voice analysis, distinctive features, or phonological processes. The treatment goal as the authors report is to eliminate the misarticulated patterns and teach the missing normal patterns. Correct production of the sound patterns should be realised in words, in sentences and eventually in spontaneous speech. Whatever the selection of the sound patterns and the method to teach them, contrast between the normal, the misarticulated, and the produced sound pattern is always important so the client becomes aware of differences in various articulatory configurations.

1.6.3 Visual reinforcement

Increasing, strengthening, and finally sustaining the frequency of correct and accurate responses of the client is one of the most crucial elements for effective treatment of misarticulations and an integral part of every modern treatment program. As it is often the case and practise has shown, this is highly related to an appropriate form of reinforcement. Feedback in general can be considered as a, *"reinforcer when it increases the rate of response or some quality of it"*, **[41]**. It is often the case that if a reward follows the response to a stimulus then the response becomes more probable. This is best met with 'instrumental reward conditioning' once a method for evaluating client's responses has been established.

1.7 Our theoretical approach to teaching speech segments

Both the broad description of the cases we are going to examine in this thesis from the linguistic perspective, *AM* and *FAD* **(§1.3.2)**, and the principle of visual contrast **(§1.6.2)** raise questions about the segmental nature of speech and the internal structure of phonological representation. According to some the access to this internal representation consists of distinctive features, according to others it consists of phonemes or allophones; and according to yet others, it consists of syllables or words. Sendlmeier, **[94]**, argues that the type of representation depends on the type of task, the context of perception, the speaking rate and/or the complexity of the stimuli. He states that the listener may *"interchangeably focus on different kinds of representation while solving one particular task. Thus, a listener can switch to single sounds or even distinctive features when discriminating, for example, minimal pairs or difficult words such as proper names, words of a foreign language or pseudowords"*. Stackhouse and Wells, **[96],** agree by saying that *"phonological representations are considered to have an internal hierarchical structure, in which units 'higher up' are at least as important as units lower down"*. In OLT we try to follow this approach, by modelling and creating appropriate targets (**§6.1**) and stimuli with the aim of allowing the speaker to focus interchangeably on either phonemes/allophones, distinctive features such as voicing or lingual stricture, and small vowel-consonant syllables.

1.8 Critical problems with non-instrumental articulation treatment

In general, although speech training without the use of any technological aid is commonly practised nowadays with some success, it does not answer successfully the questions we raised in §**1.1** because it suffers from serious drawbacks such as :

(a) Insufficient motivational impact and reinforcement.

It is very difficult to find appropriate stimuli and teaching strategies that will be both effective and attractive so that the desired response from the client can be evoked. If not, this usually results in limited concentration and performance of the speech drills, especially in the case of children.

(b) No immediate linking and response between the acoustics and the articulation.

Perhaps one of the most critical problems is that directions and instruction have to be delivered at the same time production occurs, compensating for the natural real-time auditory perception mechanisms. In non-instrumental speech training the teacher delivers and consequently the student receives feedback with some delay after the productions have occurred. The aim is to make student self-aware so that s/he can scan ahead in his/her intended utterances to anticipate misarticulations before they occur.

(c) Evaluation of articulation is based mostly on the acoustics.

Usually treatment programmes focus on the feedback the student receives. It is equally important what feedback the teacher receives. In *FAD* and *AM* cases the teacher like the student can hardly see the dynamic movements of extremely important articulators like the tongue, and relate them to what is perceived. Therefore it is quite probable that the teacher may misjudge the attempts of the student and the student may be misled because of wrong evaluation.

(d) What properties of the supraglottal articulators to alter and how

Effective communication skills and delivery of detailed instructions on how to make the client realise and report the errors occuring during speech production is perhaps the most demanding task of the teacher. Complexity increases if we also consider the directions the teacher should deliver to teach the client how to reach the target production. Even if the teacher possesses phonetics and phonological knowledge and expertise about the speech production mechanisms, which is assumed to be the case only for therapists, it will be hard for the client to understand.

(e) Practice does not continue after the speech training session.

Unless the students keep practising the new skills acquired after the end of speech training session or after treatment is finished, it is highly unlikely that they will maintain progress in error correction.

(f) Requires too much time and patience from both the teacher and the students.

All reasons (a) - (e) contribute to excess time consumption.

It is on these grounds that a computer-based system should assist the role of the instructor in speech training. As we shall see in the rest of the thesis, the computer-based visual feedback we provide; attempts to address and overcome problems (a-d), while as a side effect it can be proven beneficial for (e) and (f).

1.9 Visual feedback instrumental techniques

As we have already discussed in §**1.4**, visual monitoring of articulators like lips, jaw and even tongue in limited cases has been an augmented form of feedback in non-instrumental speech training. But we also mentioned in §**1.8-(c)**, **(d)**, that it is very hard for both the teacher and the student to see and monitor dynamically the movements of important articulators, relating them to the acoustics of speech production. This general problem of visualising the articulation and providing effective feedback through the visual modality has been addressed by several instrumental techniques (§**2.3)**.

These have quite often been incorporated in a computer-based speech training system. **Figure 1-4** describes the extra audio-visual feedback the learner and the teacher receives at the same time from a computer speech training aid. Normally the teacher perceives audio and visual cues from the training of the client and the client receives instructions and evaluation from the therapist. Therefore effective audio-visual feedback from the computer simply aims to couple the feedback of the teacher to the learner and the teacher's monitoring abilities to the client's performance.

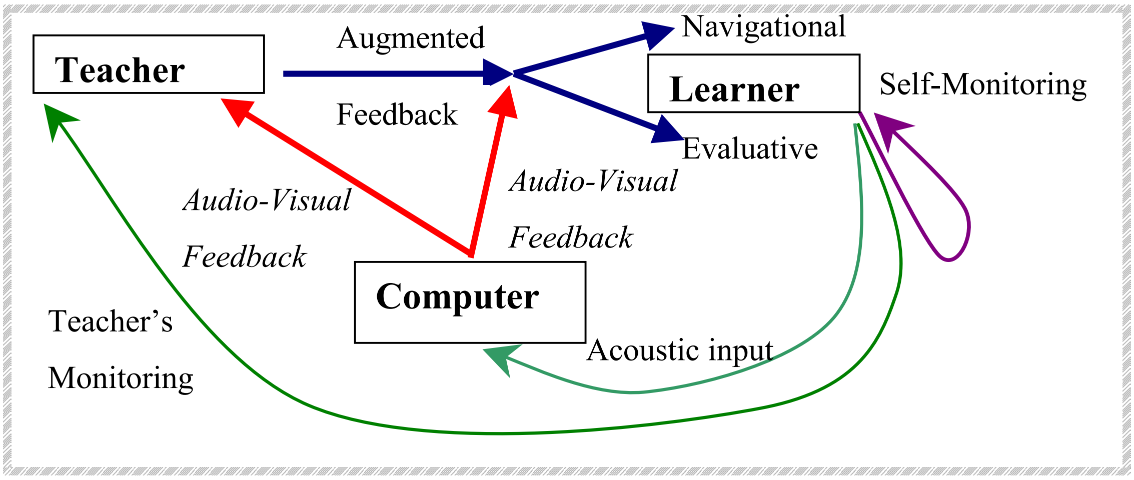


Figure 1-4 *The feedback relationship between Computer, Teacher and Learner*

On the other hand it is easier for the client to become aware of the misarticulations and try to correct them by relating and strengthening the link between speech perception and speech production if he/she attends to the visual cues provided and the temporal equivalent link with the acoustics of speech production. Clark and Yallop observe that "phonological description has always tended to be articulatory in orientation, for the obvious reason that the gestures and settings of articulatory organs are more easily observed than sound waves", **[15]**. It is perhaps these observations that motivated many researchers to develop visual speech training aids that attempt to visualise the place/manner of articulation through techniques similar to biofeedback (§**1.6.1**). The pioneers of computer based speech training systems like HARP and Video Voice (§**2.3.2F,** §**2.3.2E**) were based on the formant mapping method (§**3.1.1**). In addition, recalling the psycholinguistic explanation of *FAD* and *AM* we notice that the motor program which is dependent on the phonological representation (**Figure 1-2**) needs to be corrected. But according to Stackhouse and Wells, **[96]**, augmented feedback such as distinctive visual information is incorporated into the phonological representation, for example, *"listeners with hearing impairment are particularly dependent on such visual cues"*. Therefore by correcting or strengthening the phonological representation we have an immediate effect on the correction of the motor program.

There have also been attempts to provide other forms of feedback from instrumental techniques ; but there is a poor temporal resolution for senses other than hearing and vision, in particular the sense of touch **[86]**. Many of problems appear because of the different types of technical visual displays available and because of how this technology is applied and fits into the existing speech training programmes. Certainly in many cases one or more of the critical problems referred to in (§**1.8**) are either partially solved or not solved, and others have been added. A detailed discussion is given in **(§2.3)**.

1.10 The challenge and the principal aims of the thesis

In this thesis we develop and present a full implementation and application of OLT visual feedback based on a piece of software which we call OLTK (Optical Logo-Therapy Toolkit) that uses speech signal processing, statistical, connectionist, and software engineering techniques. OLTK has been designed to include appropriate interactive displays for both the teacher and the student. The principal aims are :

(a) Provide the teacher and the student with real-time, audio-visual feedback based solely on acoustic input from a microphone.

(b) Visualise the contrast between different articulatory configurations.

(c) Visualise the variation of certain articulatory features.

(d) Motivate and reinforce the learning with appropriate visual stimuli.

(e) Evaluate the attempts of the client using scoring metrics.

The key idea in the design of the software toolkit is called 'phonetic maps' which we create by training a neural network to associate acoustic input vectors with points in a two-dimensional space.

1.11 Structure of the thesis

Now, in order for the reader to get a general view of the material that appears in the rest of this thesis we present a brief summary of the contents for each one of the rest individual chapters:

Chapter 2 formulates the requirements for effective visual feedback in speech training. We review computer-based and other visual displays in the light of these requirements.

Chapter 3 focuses on two-dimensional speech training displays. We describe the techniques which are commonly used to portray a 2D visual representation of the speech signal and their advantages and disadvantages. We summarise the critical design aspects for 2D speech training displays.

Chapter 4 presents the OLT toolkit, OLTK, from a software engineering perspective. It explains the metrics used to evaluate speech quality and the technique for mapping from acoustics to phonetic map.

Chapter 5 presents the OLTK graphical user interface and explains the functionality that OLTK supplies to the teacher and the client. We emphasise OLTK 'sensitivity' control and real-time evaluation features.

Chapter 6 reports on the application of OLT in treating functional articulatory disorders and accent modification. We describe the experimental methodology and the results. We discuss problems which surfaced in the speech training sessions.

Chapter 7 reviews the OLT method and its effectiveness for speech training, basing the evaluation on the requirements set out in Chapter 2. Separately, we evaluate the software toolkit OLTK. The thesis ends with future plans and conclusions.

Chapter 3 and *Chapter 4* can be skipped by readers without computer and speech technology expertise.

Chapter 2

*It was our sorry case that caused the **Logos** to come down, our transgression that called out His love for us, so that He made haste to help us and to appear among us. It is we who were the cause of His taking human form, and for our salvation that in His great love He was both born and manifested in a human body(1:4). Man, who was created in God's image and in his possession of reason reflected the very **Logos** Himself, was disappearing, and the work of God was being undone (2:6). For He alone, being **Logos** of the Father and above all, was in consequence both able to recreate all, and worthy to suffer on behalf of all and to be an ambassador for all with the Father (2:7).*

St. Athanasius-On the incarnation (Translation-C.S.Lewis)

Visual Displays in Speech Training

2.1 Historical Background

Nowadays computers and other technical devices have enhanced a lot the speech training abilities by visualising the articulation process through various technical displays. Originally many of these devices have been used to provide visual feedback for the speech of hearing impaired clients, but later on the use has been broadened to other categories of speech disordered clients.

Spectrographic analysis was the first attempt to relate what were already known as articulatory properties to what are now known as acoustic or spectral properties, **[78]**. Spectrography can be considered as a type of a detailed acoustic analysis and as such can be used to detect phonological/phonetic disorders and analyse distortions, substitutions and omissions, **[23]**. According to Maki, **[64]**, the visual characteristics of speech spectrographic displays suggest application of the instrument in learning the articulatory features such as voicing, manner and place of production. It also involves instruction associating the acoustic features of speech with the appropriate articulatory movements. Although speech spectrography is quite a useful method in speech training, a significant amount of time needs to be spent on teaching students how to read and interpret spectrograms. As Maki reports, "*Display characteristics do not provide distinctive contrasts for all phonetic differences. And finally it must be considered that an analytic, visual approach may not be appropriate for all learners*", **[64]**. Therefore we can conclude that perhaps spectrograms are great for phoneticians but not for clients.

Clearly speech training with this type of visual feedback suffers from problems which we already mentioned in §**1.8, (a), (d), (e),** and **(f)**. Obviously the solution to providing proper visual feedback for the treatment of *FAD* and *AM* cases depends upon certain criteria and design specifications of the visual displays.

2.2 Practical considerations for effective visual feedback

It has already been mentioned in §**1.9** that visual feedback provided by various technical devices should be fully integrated in current non-instrumental articulation treatment programmes. This is necessary in order to address the critical problems that we discussed in §**1.8**. Now we are going to set down the most influential and important factors, so that visual feedback provided from the visual displays of such devices can be effective for both the teacher and the student. Speech training displays should address the following criteria :

A. Time between response and visual feedback

It is critical whether the feedback is immediate or there is some delay between the time when speech is produced and the display of visual information for the production, **[73]**. Watson and Kewley-Port, **[99]**, state that : *"neuromotor responses can be associated more quickly with the production of specific speech sounds, if the loop between those responses and the feedback is as short and direct as possible"*. We want feedback to be immediate, so that the temporal equivalence of what is being spoken and what appears on the screen is clear. We also need to be able to reinforce this link by replaying recorded sounds in parallel with their visual display. Nevertheless, delayed feedback or gradually reducing the dependence on feedback can be useful for testing whether the student learned the specific speech skills that s/he was taught.

B. The level and amount of detail extracted

One of the issues often raised in the process of designing speech training displays is the effect of isolating specific speech characteristics. For example, we can have displays that focus only on one feature of articulation like loudness or voicing, and we can have displays with a high level of detail as in a spectrogram or in a two-dimensional vowel matrix. The single-feature displays form a useful support tool for the teachers allowing them to draw attention to features of interest but cannot be related easily with the overall process of articulation. On the contrary high-detail feedback as a combination of more than one physical properties of the articulators can be much more effective during speech training, **[99]**. A disadvantage of complicated displays is that they can be confusing and incomprehensible. Another problem here is "*the dilemma to provide specific speech training in a particular voice or phonetic characteristic without inadvertently giving experience and reinforcement for inadequate or abnormal performance of another characteristic*", **[9]**. Ideally, the display should provide sufficient detail to reinforce the learning with a reasonable level of abstraction. We are focusing in the segmental phonetic characteristics of articulation and we examine features like lingual stricture, and labial setting.

C. Characteristics of visual representation

On the other hand the visual feedback has to be closely related to the specific speech attributes taught. Therefore the visual representation of the display has to be natural, logical and easily comprehensible, **[73]**. Intensity, for example, can be associated with the size of an object that becomes larger as a sound becomes louder or vice versa.

For the *FAD* and *AM* cases we examine it is necessary for the feedback to strengthen the association between articulation and acoustics, so that consistent changes in place and manner of articulation should produce correspondingly consistent changes in the display. It should be natural for the client to make the link between changes in her/his articulation and changes in the display, without the need for specialist speech knowledge. Therefore spectrograms and the like should be made available for use by the teacher, but should not form the basis for the feedback which the client sees.

Especially in the case of young children, there are limited concentration abilities and poor manual dexterity in interaction with a system, so the display has to be designed to be as simple and attractive as possible.

D. Visual instructions and directions on how to reach the target production

Ideally, there should be visual indications, instructions and directions shown on the speech display about how to correct the error and reach the target production.

Navigational feedback has more to do with the association between articulation and acoustics and how these are visually represented on the display. The trainee not only observes, but systematically tries to control consistently the events on the visual display, aiming to produce specific visual patterns that correlate with the correct position of his/her articulators. Clearly, in our case the animation of visual events on the display should be related directly to the dynamic movement of the supraglottal articulators. Therefore, there should be a visual representation of particular gestures associated with the acoustics of the speech signal and also a correlation between the variation of specific articulatory features with visual qualitative changes on the display. That form of feedback we call navigational-qualitative.

E. Meaningful contrasting models

One of the basic functions of the speech display is to exhibit the characteristics of speech in such a way that different gestures are clearly and consistently displayed in an easily interpreted form. Clients should be able to compare their articulation with a model from utterances of the instructor, or with a model from a database of other speakers' recordings, or with a model from their previous past efforts, **[99], [73]**. Since we are interested in specific articulatory configurations the models created from acoustics should reflect the acoustic variation, and limits on the movement of the articulators, to shape a particular configuration.

F. A metric to measure speech quality

Requirements {C}, {D} and {E} imply that the judgement on the articulation by the client is based purely on qualitative characteristics of the feedback. The client simply tries to see certain visual patterns and animated events on the display, and qualitatively assess his/her attempt.

Normally in traditional speech training after each attempt, the teacher judges the quality of speech production and provides evaluative feedback with words and phrases like "You can do better", "Good", "Excellent", "Super", etc. The same tactic can be followed with the qualitative feedback provided by a visual display. According to Katz, **[50]**, *"the patient's response (whether keys pressed, buttons pushed, or controlled moved) should be matched against a predetermined target response or range of* *responses. At the completion of the task, a screen should show the patience (and clinician) performance scores for the session in table or graph format"*. Once a metric is defined and computed to represent and measure speech quality, we can have a scale to assess the articulation of the client. This metric is usually built as an estimate of the similarity between the student's attempt and the sound model. The metric can also be based on the acoustics to articulatory visual transformation or relationship. Whatever the case, this additional approach for providing feedback to the client can be characterised as evaluative-quantitative.

G. Motivational impact and reinforcement of visual feedback

This aspect plays a key role in effective treatment because, a lot depends on whether the visual feedback provided is attractive enough to stimulate and encourage the student to attain the desired response, **[73]**. This criterion is related to the traditional behavioural conditioned learning of Van Riper. Today most computer speech training displays are enhanced using a series of video-games formats, in which the goodness score from a certain evaluation of speech production, requirement {F}, determines the size and/or the location of various graphic objects. Motivation and reinforcement is also strengthened from other requirements {C}, {D}, {E}, that are all prerequisite. Katz reports, **[50]**, that *"Stimuli should be displayed in a consistent, straightforward, and uncluttered format"*. This is a particularly desirable feature in case the clients are children in order to hold their attention and interest.

H. Flexibility and suitability

(i) Adaptable to specific FAD or AM problems

Rather than attempting a single mapping from any speech input onto a fixed display, we wish to tailor the speech display to suit specific *FAD* or *AM* problems. In this way we can effectively focus on training of particular problematic segmental characteristics of student's articulation.

(ii) Adjustable to the varying performance of the student

In general the system needs to be flexible in the training, because speakers vary widely in their needs and in their ability to reach target production, **[73]**. That depends on the level of disorder and the various personal characteristics of the student, such as cognitive ability, dialect, age, gender, and vocal tract parameters. The teacher should be given the option to tailor the displays and adjust accordingly the visual feedback depending on whether the target production is attainable or not. This tactic guarantee that the student will not become discouraged and frustrated, and will not be challenged to achieve a goal that is beyond his or her capability, **[99], [88]**, but instead the potential to reach the target production progressively increases.

I. Accuracy of the visual feedback

In order to talk about the accuracy of the visual feedback we have to distinguish between the two different kinds of feedback we have defined, the navigational-qualitative and the evaluative-quantitative one (**§1.4**).

(i) Accuracy of the navigational feedback

The accuracy of navigational feedback is based on requirements {C} and {D}. According to this, consistent changes of supraglottal articulators should produce correspondingly consistent changes on the display.

(ii) Accuracy of the evaluative feedback.

Accuracy of the evaluation depends on the basis of the metric. The highest level of sophistication is achieved by applying speech recognition techniques to evaluate the speech quality of an utterance in comparison with a model. This technique runs the risk that recognition results may not correlate transparently with changes in articulation. The judgement of the machine may well differ from that of the teacher, and evaluation may not be objective and reliable **[73], [99]**. Kewley-Port and Watson report, **[54]**, that *"at this time the only suitable metric of speech quality, especially for disordered speech is the judgement of the human listener. Therefore, this validation requires a correlation between the quality metric calculated and the quality ratings of these same utterances by human listeners"*. Evaluative feedback has to be extremely accurate because the teacher cannot correct inadequate productions that the system itself judges as adequate and evaluates them with a high score, **[81], [62]**.

Special attention has to be paid to the accuracy of the feedback. If the feedback is not accurately displayed, it will allow the speaker to continue making errors ; in the worst case it is possible to reinforce bad productions and that could be extremely damaging to the development of speech production. Moreover wrong feedback can also mislead the teacher on his/her judgement. This new requirement has to satisfy and is also related with {Hii} as the quality of the feedback depends on the performance of the student.

A synopsis of the above requirements gives us the following descriptive characteristics for what we will call effective visual feedback (EVF), provided by any speech training display.

{A}	**Immediate**
{B}	**Specific, Focal**
{C}	**Consistent, Natural, Logical, Comprehensible**
{D}	**Systematic, Dynamic, Navigational, Instructional**
{E}	**Contrasting**
{F}	**Evaluative**
{G}	**Motivational (Attractive, Interesting)**
{H}	**Suitable(Adaptable, Adjustable), Flexible**
{ I }	**Accurate (Reliable, Objective)**

Table 2-1 *List of requirements for effective visual feedback (EVF)*

2.3 Review of modern visual displays used in speech training.

Nowadays because of our modern technological advances, several instrumental techniques have contributed significantly in the analysis of articulatory mechanisms and the visualisation of the speech events. Nevertheless, the use of these techniques in the treatment of *FAD* and *AM* cases appear to be disputable and ambiguous. We would briefly present what are the most representative of these techniques related to FAD and AM and how they comply with our requirements in **Table 2-1**.

2.3.1 Acoustic vs Electrophysiological visual displays.

So far we have not said anything about the source of feedback displayed. Here we can distinguish between acoustic, electrophysiological or mixed sources. In the first category, visual feedback is obtained from acoustic analysis of the speech signal through a microphone. As a result, acoustically-based visual displays are relatively simple and cheap and have in general been the main choice of commercially developed speech training systems in the past **[8]**. The main drawback here is that a number of speech parameters such as nasalisation, or consonant articulation, cannot easily be obtained from the acoustics only. For that reason it is quite common to supplement the acoustic information with other physiological measurements.

The best known electrophysiological techniques are electropalatography (EPG), **[26]**, **[34]**, glossometry **[27]**, ultrasonic, **[51]**, electromagnetic articulography (EMA), **[97]**, **[74]**, and magnetic resonance imaging (MRI), **[5]**. Some of these techniques, e.g. (MRI), have been used mostly for medical diagnosis. The disadvantages of physiologically-based systems are considerable, however. The equipment required to obtain the measurements is often delicate, expensive and difficult for non-experts to adjust and use **[6]**. Moreover one has to consider that the techniques are relatively invasive when compared with acoustic measurements, especially for children. According to Moll's principle, **[69]**, there should be as few restraints as possible on normal speech activity so that interference with normal and natural articulation is as benign as possible and with a minimum of invasive interruption. Certainly these techniques increase our monitoring abilities of dynamic movement of articulation and increase our knowledge, particularly on kinematic information of the supraglottal articulators. Nevertheless many of the criteria for *EVF* listed in **Table 2-1** may not be satisfied. To exemplify this we discuss *EPG* feedback in the next section.

A. Electropalatography (EPG)

The specific *FAD* case that we examine in this thesis, has been studied with electropalatography. *EPG* has the advantage of effectively making a real-time, {A}, direct measurement of the lingua-palatal contact. The visual feedback provided can certainly be judged as natural and logical, {C}, and certainly contrasting, {E}, but very specific, focal {B}. However Hardcastle reports, **[35]**, that *"The technique records the location of the contact only; there is no direct information on proximity of the tongue to the palate; nor is there any way of inferring directly which part of the tongue is producing the particular contact pattern"*. Therefore although differences between a /s/

and a /ʃ/ can be shown visually as patterns of contact between the tongue and the hard palate (**Figure 2-1**) the proximity of reaching those places of articulation cannot be adequately captured from the palette. Here one should also consider the problem of co-articulation and how different and approximate are the articulatory positions from those spoken in isolation.

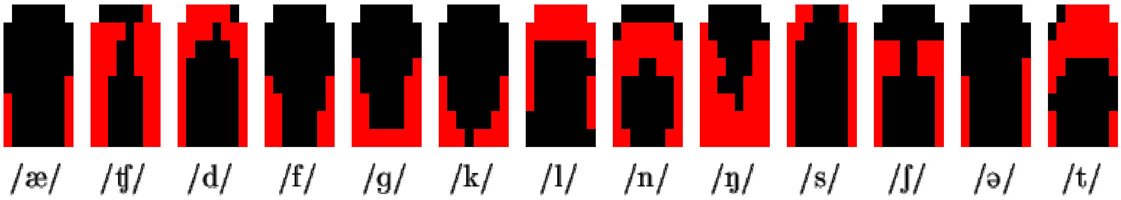


Figure 2-1 *EPG frames. (Captured from www pages of Miguel A. Carreira Perpipan)*

What is also due to the fact that as Hardcastle reports *"another difficulty in interpreting the contact patterns in terms of tongue movement arises from the fact that the electrodes are discrete points ....."*, **[35]**. **Figure 2-2** demonstrates that observation. It is quite hard for such display to make clear detailed differences of the place of articulation. That makes difficulties for the therapist, and of course the client, to interpret the patterns and deduce the exact place of articulation. Moreover the voicing, nasality and labial setting that often play a critical role in articulation cannot be distinguished with *EPG*, **[26]**. All these aspects raise questions about how comprehensible, consistent, and accurate the displays from *EPG* can be, {C}, {I}.

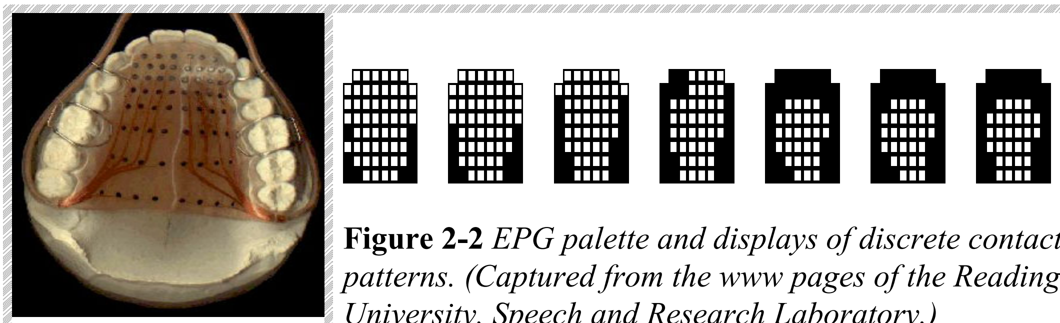


Figure 2-2 *EPG palette and displays of discrete contact patterns. (Captured from the www pages of the Reading University, Speech and Research Laboratory.)*

Further on, Hardcastle notices that *"it is obviously important that the palate should interfere as little as possible with normal speech production"*. This highlights the problem of invasiveness. Another important problem is that although *EPG* displays can be used to capture the tongue dynamics, shown as a series of discrete stages (**Figure**

2-2) the contrasting and instructional requirements, {D}, {E}, can hardly be met for segments of speech but only for teaching single articulatory positions. According to requirement {F}, *EPG* display itself is not evaluative. There can also be a strong argument about the motivational impact of *EPG*, not to mention the absence of reinforcement from such a display especially with children who do not feel particularly comfortable with invasive techniques {G}. Moreover it cannot be considered adjustable to the varying performance of the client. You can freeze, capture and display one or more interesting frames or sequences of frames but it is very hard to do a comparison in real-time with all of them appearing concurrently on the display. Finally because of the design specifications it is non-flexible, as a palate has to be constructed for each client, and certainly because of that it is expensive. For those last reasons practise cannot be continued at home.

Despite the limitations and the problems of *EPG* a large body of literature exists together with a record of success in clinical remediation, **[16], [29], [34], [35], [46]**. Nevertheless, it seems that the cases examined are mostly related to teaching positions for alveolar and velar stops, sibilant fricatives and affricates, and in some of them the post-therapy review is neglected, or not reported sufficiently.

2.3.2 <u>Acoustic computer-based visual displays</u>

A. Spectrography

With the rapid development of micro-computers spectrograms can now be produced in real-time, {A}, but it is difficult to read the displays and the relationship of visual spectral qualities with the articulatory features is hard to interpret, {B}, {C}. There can be contrast, {E}, but it is not clear how to vary the articulators in order to visualise a certain desirable pattern on the spectrogram {D}. It is a qualitative approach, and the accuracy depends on the spectrographic analysis, {I}. Moreover, it does not provide reinforcement and has limited motivational impact, {G}. Nevertheless many modern speech training systems incorporate such visual displays, as sometimes it is useful for the expert to detect certain spectral characteristics of speech and combine them with other visual displays.

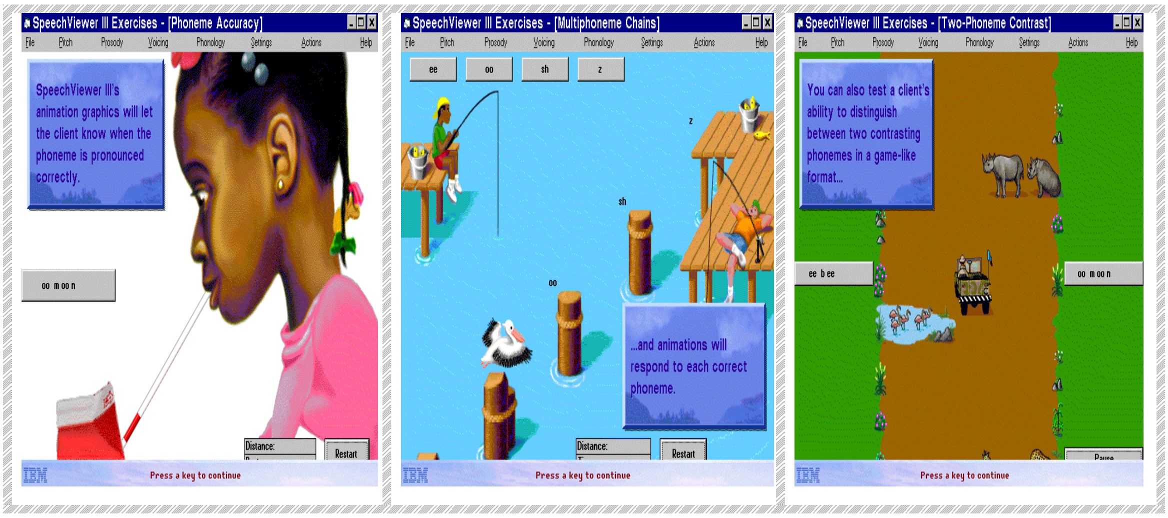
B. Vocal track visual displays

There have been many attempts to display accurately information on the shape of the speaker's vocal tract as an aid in the training of vowel production. Two of them, the 'Computer Vowel Trainer', **[13]**, and the 'Vocal Tract Area Display', **[90]**, use linear prediction analysis (LPC), to generate an approximate vocal tract shape for sustained vowel articulations in real-time, {A}. The display can be frozen at any time and stored to provide a static target for imitation by the speaker, {E}. If the same vocal-tract shape as the model is achieved, some form of reinforcing feedback is given to the student, for example a cartoon teddy bear which smiled, {G}. The accuracy of the match is controlled by a variable threshold setting {H}.

There are problems associated with these attempts to use vocal tract area functions, in particular the production of "impossible" vocal tract shapes owing to the inadequacies of LPC {I}. There are modern and far more accurate techniques which either simulate or visualise the vocal tract but these are not dynamic or not in real-time, e.g. MRI, {A}. The system can be used only in the training of the vowels in isolation ; thus it has limited flexibility, {H}. The display can also be considered too detailed and technical, {B}, {C}, non motivative, {G} and there are no visual indications and directions as to how to produce the target, {D}. Finally no evaluative form of feedback is provided {F}.

C. IBM SpeechViewer

This can be considered one of the best speech training systems nowadays. It was released as a commercial product in 1989, **[7]**. IBM SpeechViewer is an acoustic system that provides real-time visual and auditory feedback {A}. We are going to examine the sustained phoneme and speech segment production visual displays of the system. The skill-building strategy of IBM speech viewer employs a "goal-oriented" methodology. This provides motivation in the form of bright and interesting graphic displays, movement toward a goal, and positive reinforcement when the task is complete and successful {G}. Nevertheless that form of feedback is mostly evaluative and is based on a metric for spectral comparison with a phone model {F}. The triggering of events on the screen and the locations of various objects depends on the goodness-of-fit score and a threshold setting (**Figure 2-3**) {H}. That reveals the weakness of the system in catching the dynamic real-time changes of specific articulators and to relate them systematically with the visual display {D}. Therefore contrast can be made only between the phone models and the production of the student {E}.


SpeechViewer III Exercises - [Phoneme Accuracy]
File Pitch Prosody Voicing Phonology Settings Actions Help
oo moon
Distance:
Restart
IBM
Press a key to continue
SpeechViewer III Exercises - [Multiphoneme Chains]
ee
oo
sh
z
SpeechViewer III Exercises - [Two-Phoneme Contrast]
ee bee
oo moon
Pause

SpeechViewer III's animation graphics will let the client know when the phoneme is pronounced correctly.

...and animations will respond to each correct phoneme.

You can also test a client's ability to distinguish between two contrasting phonemes in a game-like format...

Figure 2-3 *The "goal-oriented" strategy of IBM speech viewer.*
(Figures captured from a demo version of IBM Speech Viewer III)

Since the models are always related to positive reinforcements through the games it does not make sense to include models that represent abnormal gestures. This is an important deficiency, as the student is expected to repeat previous mistakes in case he/she forgets the new skills of production acquired ; and in that case s/he cannot contrast with the wrong configuration. Visual feedback can be assumed accurate as long as the models have been built properly {I}.

The SpeechViewer has undergone a number of evaluations. In one of them, **[81]**, a number of practical problems with the aid became evident : inability to sustain children's attention over multiple sessions, inaccurate feedback with low-intensity, hypernasal productions, and the aid proved to be not too instructional and in some cases cognitively demanding. More recently Öster, **[73]**, agrees that Speech Viewer should be viewed as a supplementary tool for occasional use in speech therapy.

D. Indiana Speech Training Aid (ISTRA)

ISTRA, **[53], [55], [99]**, follows the same principles as IBM speech viewer but with a commercial speaker-dependent isolated word recognition board. The system processes acoustical input at near real-time, {A}, and provides a numerical score showing the match between the user's attempt and a stored template for the word in question, {E}, {F}. This score is used to supply the user with visual feedback though animated games. Templates can be derived from the speaker's best previous productions, {H}. The speech displays offered for training are either simple bar graphs or target-oriented games that provide positive reinforcing feedback if the attempt reaches or exceeds a certain threshold, {G}, {H}. Perhaps the only basic difference from the IBM speech viewer is that the system is better tuned to the recognition and evaluation of words rather than single phone contrasts.

ISTRA has been evaluated with many trials and has demonstrated significant progress on the speech production of the articulatory disordered clients, **[55]**.

E. The Video Voice Speech Training System

This is a commercial system marketed by Microvideo that analyses the speech from the acoustic waveform and extracts information on fundamental frequency, loudness and spectral quality. Speech parameters can be represented both in a parametric variation over time and using video games according to a numerical score that represents matching of a certain target for each attempt of the student, {F}. Although the feedback provided is in real-time, {A}, there are problems with the accuracy and the basic of metrics for the scoring, {I}, **[25]**.

Video Voice's "F2 vs. F1" matrix display (**Figure 2-4**) helps train correct vowel production and word articulation. Isolated vowels appear in general regions of the screen, and words have shapes determined by their phonemic components. In practise both the original model and the trial voice pattern are held on the screen for visual contrast, {E}. Some consonants are also visible on the display. In some other format of the display, called 'eat the dots' the object is to erase the model through repeated production of the target. That game stimulates many repetitions of the target utterance and helps the clients to find and maintain appropriate articulator positions, {G}. All models are instructor-defined, so s/he can determine the program content and select appropriate training targets for each individual in the caseload, {H}. We are going to discuss the technical details of this kind of display in the next chapter.

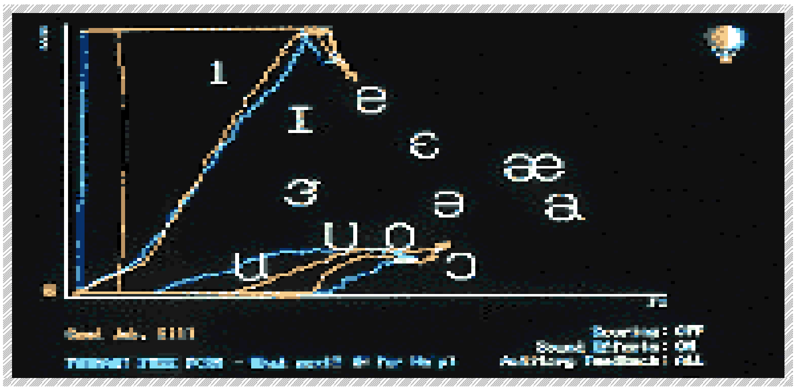


Figure 2-4 *Video Voice "F2 vs. F1" matrix display.*
(Figure scanned from an advertising leaflet of Video Voice)

Users found that formant displays are difficult to understand, {C}, and had to be interpreted by the clinician. Furthermore, since the spectral information was limited to F1, F2 formants, system was suited to teaching vowels rather than consonants, **[25]**. Despite its limitations Video Voice system has been evaluated, and was judged to be a useful tool for therapists.

F. Speech Rehabilitation Speech Training Aid for the Hearing Impaired (HARP)

HARP, **[87]**, **[89]**, uses similar displays to Video Voice to teach vowel articulation in real-time, {A}, {B}. The vowel analysis uses a method of cross-speaker normalisation to obtain a two-dimensional representation of vowel quality from the first two formants and the fundamental frequency.

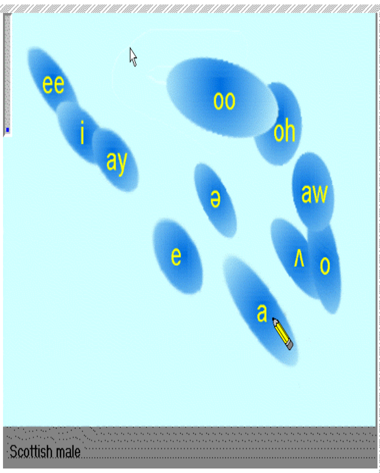


Figure 2-5 *HARP vowel display.*

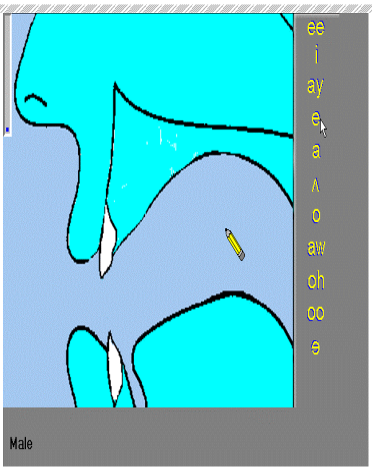


Figure 2-6 *HARP vocal tract display*

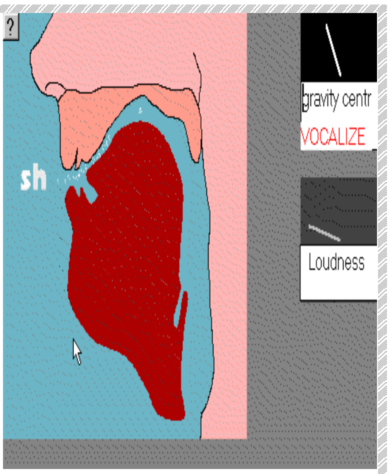


Figure 2-7 *HARP vocal tract fricative production*

All figures captured from a demo version of HARP

These two acoustic dimensions (F2-F1, F2-F0 differences) correspond approximately to the articulatory dimensions of tongue frontness-backness and tongue height respectively (**Figure 2-5**) {C}, {D}, {E}. It is also possible for the teacher to move the targets in positions not too difficult to reach or leave on the display only the targets of the vowel of interest. Alternatively a vocal tract display can provide real-time visual feedback about the position of the tip of the tongue **(Figure 2-6)**. The symbols on the right side of the display can be dragged inside the vocal track as reference points. Finally, production of the palato-alveolar sibilant fricatives can be controlled with a real-time animation of the cross section of the vocal tract (**Figure 2-7**) where the tongue shape is shown approximately. There have been no officially reported evaluations of the system in speech training.

G. Visual Speech Apparatus (VSA)

The Vowel Corrector, **[80],** is one of the precursors of VSA. It processes acoustic data using linear discriminant analysis to plot vowels in a 2D display and was designed specifically to offer simplified visual feedback on vowel production. Evaluations were conducted and the system was found to be motivating and reinforcing, {G}. The main problem as authors report, **[80]**, was that *"the training of spectrally similar vowels was sometimes affected by the relatively large overlap between different vowels."*. We discuss similar kind of displays and problems in Chapter 3.

At about the same period Zahorian and Venkat developed their own computer-based aid for teaching vowel production known as the Vowel Articulation Training Aid (VATA), **[105]**. The system is based on their mean square error coordinate transformation (MSECT) algorithm, a general linear transformation of the acoustic data into two perceptual dimensions, **[104]**. This derives a 2D visualisation of vowel articulation which is intended to be reliable, easy to interpret, and directly related to changes in vowel quality. We will analyse and describe the technique in more details in **§3.1.2IIA**.

Arends and Povel combined the methods used in the above two aids to construct 2D vowel and consonant displays, {B}, used in the Visual Speech Apparatus (VSA), [3]. They use two classification methods the 'direct' and the 'hierarchical' one. The 'direct classification' is solely based on the MSECT algorithm of Zahorian (§**3.1.2IIA**). The 'hierarchical classification' uses a combination of linear discriminant analysis and the MSECT algorithm. In order to increase the discriminative capacity they limit the number of phoneme classes so that the between-class variance is lowered and the ratio of the between-class variance and the within-class variance is increased. Then they use the MSECT algorithm to obtain weights that determine the distance of an input phoneme relative to all phonemes in the reference set and the distance relative to all possible pairs of phonemes included in the reference set, {F}.

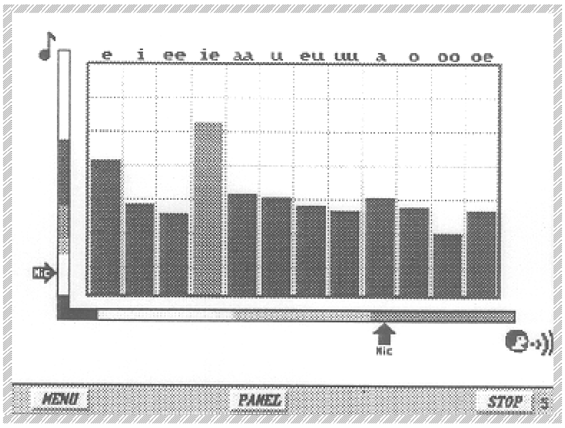


Figure 2-8 *A vowel training display based on the hierarchical classification of VSA that shows loudness (vertical), pitch (horizontal), and classification score (bar height). (Scanned from* **[3])**

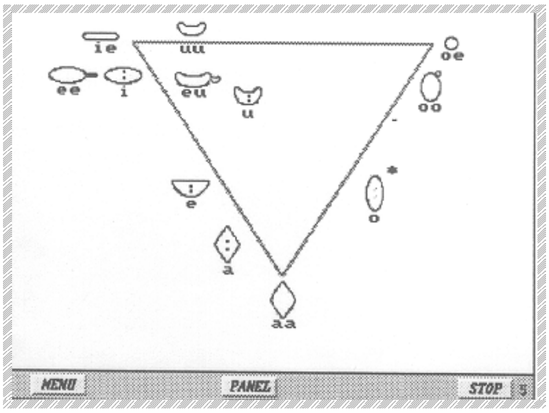


Figure 2-9 *A vowel training display based on the hierarchical classification of VSA that depicts selected phonemes in discrete positions. (Scanned from* **[3]***)*

Once weights have been calculated, classification can be operated driving the displays in **Figure 2-8** and **Figure 2-9**. The visual feedback provided is simple and attractive, {G}. Displays that have both vowel and consonant phoneme categories have not been developed in VSA. Feedback is immediate, {A}, but it can also be deferred in order to reduce the user's dependence on visual information. VSA has been evaluated thoroughly from its developers, [4], and showed that it can be applied successfully as an aid to speech therapy.

2.3.3 *Summary of the review*

The acoustic computer-based visual displays that are used for the treatment of FAD and AM cases can fall into three basic categories. First we have those that present a midsaggital cross-section model of the vocal tract and try to visualise the position of the basic supraglottal articulators (like tongue, lips, jaw) in real-time (**Table 2-2, 3**). This can be judged as a rather highly detailed and very technical feedback, addressed more to the expert than to a language instructor or student. There are a lot of problems that have to do with the accuracy of the simulation of the dynamic properties of the articulators and the computational model of the vocal tract.

	Visual Display	**{A}**	**{B}**	**{C}**	**{D}**	**{E}**	**{F}**	**{G}**	**{H}**	**{I}**
1	EPG	☒	☒	☒	□	☒	□	□	□	□
2	Spectrography	☒	□	□	□	☒	□	□	□	☒
3	Vocal track	☒	□	□	□	☒	□	☒	□	□
4	Speech Viewer	☒	☒	☒	□	☒	☒	☒	☒	□
5	ISTRA	☒	☒	☒	□	☒	☒	☒	☒	□
6	Video Voice	☒	☒	□	□	☒	□	☒	☒	□
7	HARP	☒	☒	☒	☒	☒	□	☒	☒	□
8	VSA	☒	☒	☒	□	☒	□	☒	□	□

Table 2-2 *List of requirements for (EVF) vs speech training displays.*

Then we have displays from systems that use an evaluative feedback based on a metric from spectral comparison or more sophisticated speech recognition algorithms (**Table 2-2, 4-5**). Although the feedback is in real-time, there could be a temporal separation between stimulus presentation and the subject's response. The stored templates try to model either a sustained phone articulation or an utterance but they do not reveal the dynamic articulatory characteristics of the gesture and the existing acoustic variability, nor the relationship between articulatory features and acoustics. It may be also the case that the entertaining motivational and reinforcement characteristics of the display can cause distraction to the client neglecting the awareness of the positions or smooth movement of the articulators from one configuration to another. In addition the teacher simply expects that the skills will be learned automatically once the correct target is reached by reinforcement of the good attempts. But it is possible to fall back into previous speech errors without realising what he/she is doing wrong, as there is not any visual contrast and connection between wrong and correct gestures, between wrong and correct attempts. That effect can become quite evident because of the carry-over problem. As a last point here we have to mention that much of the accuracy depends on the setting of the models and in some cases that is not an easy task for the teacher to do, plus the possibility that the models may be proven to be inadequate for the evaluation of student's segmental articulation. That becomes an even more complicated issue with the adjustment of the threshold how an incoming unknown input is accepted or rejected in comparison with a stored model.

Finally we reported on visual displays that provide a graphical method of showing where a speech sound, such as a vowel, is located in both "acoustic" and "articulatory" space (**Table 2-2, 6-8**). Although these displays are less motivative than their evaluative competitors the visual feedback provided is instructional, revealing dynamic properties of the supraglottal articulators like tongue position. They have been proven to be valuable in speech treatment and as we have seen that they are included in many modern computer-based speech training aids. That is what we consider to be a better approach than the two previous ones. This is going to be our topic of discussion in the next chapter.

Chapter 3

*And the **Logos** became flesh and dwelt among us, full of grace and truth; we have beheld His glory, glory as of the only Son from the Father.*

St. John-Gospel (1:14) (Translation-RSV)

*The **Logos** of God came in His own Person, because it was He alone, the Image of the Father Who could recreate man made after the Image (3:13).*

St. Athanasius-On the incarnation (Translation-C.S.Lewis)

Two-Dimensional Speech Training Displays

The potential application of any visual feedback display logically depends on the variety of articulatory features which can be differentiated using the visual patterns. The display should present a visualisation of the speech events and show where and how articulation has gone wrong. That will enhance the sense of the movement of the articulators in feedback through visual perception, and provide qualitative contrast between different sounds or utterances. In §**2.3.3** we highlighted a particular type of display that attempts to relate the acoustic properties of the speech signal with articulatory positions. This requires a technique to associate the acoustic vectors with points and trajectories on a 2D plane. This technique seems to be the most appropriate for implementing the 'visuomotor tracking' principle we discussed in §**1.6.1**.

3.1 2D visual representations of the acoustic speech signal

According to Povel and Arends, **[79]**, there are three important practical aspects for building a visual aid. The first is the extraction of speech parameters, the second is the transformation and mapping on visual dimensions and the third is the addition of the norm (standard, model) for comparison. Usually, the speech parameterisation provides us with a multi-dimensional feature vector; we are interested in mapping that vector on a 2D visual display. It is also very important for *FAD* and *AM* cases to consider how positions and movements on this 2D display can be correlated with articulatory gestures. There are two alternatives for building speech training displays. One is based on formants extraction, the position and bandwidth of the resonances from the transfer function of the vocal tract, and the other uses overall spectral shapes and methods for multivariate data visualisation. There are several arguments favouring the use of whole spectra, **[3]**. First, the algorithms designed to estimate the short-term spectral envelope are relatively fast compared with the speed of the rather complicated algorithms needed in formants detection. Second not all the phoneme categories can be represented sufficiently by formant patterns and third the formants of misarticulated sounds are often poorly defined.

3.1.1 Formant maps

The formants 'mapping' is based on formants extraction. As we have seen in **§2.3.2-E**, **§2.3.2-F**, it has been frequently used in speech training, particularly for vowel-like phoneme categories.

A. F0, F1, F2, F3 axis combination mapping

The most well known and traditional form of acoustic mapping based on formants is the vowel chart. The frequency distributions of the first three formants enable us to label the vowel quality and determine their similarity from the acoustics **[82]**. The chart can be obtained by plotting F2 on the horizontal axis and F1 on the vertical axis.

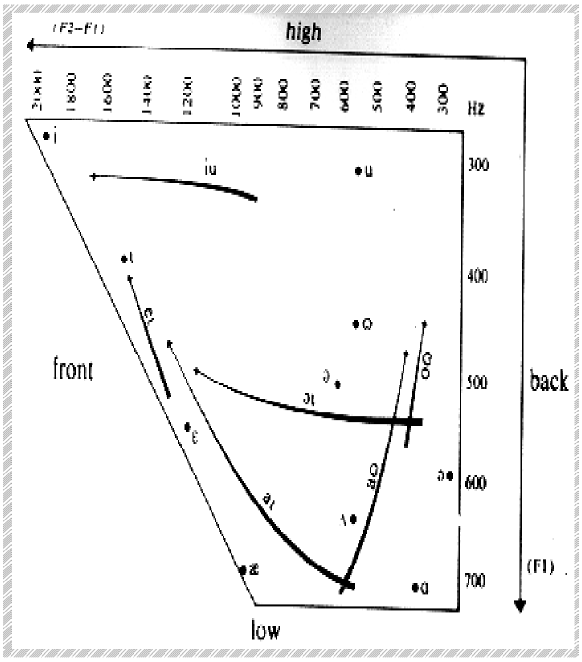


Figure 3-1 *Ladefoged's vowel chart of F1 (vertical) vs F2-F1 (horizontal). Scanned from* [58].

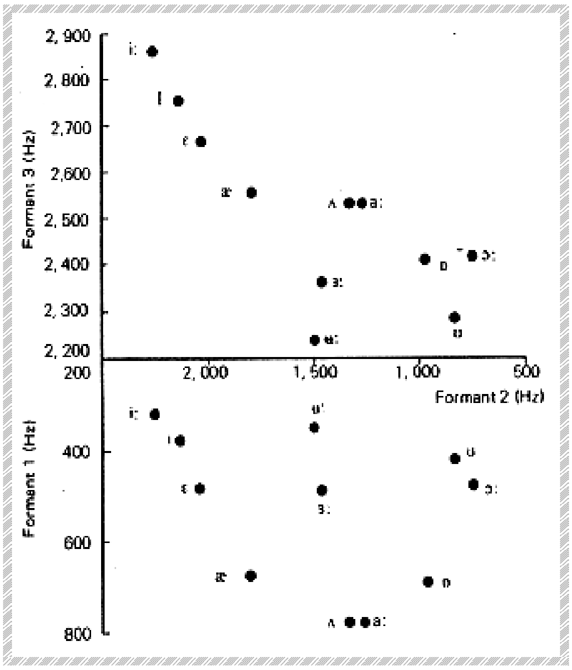


Figure 3-2 *Vowel mapping using three formants. Scanned from* [15].

Nevertheless Ladefoged replaces the F2 dimension by the difference between F2 and F1 (**Figure 3-1**). This is because F2-F1 is better related to the auditory concept of 'frontness' or 'backness' than F2 alone, while the 'height' can be sufficiently correlated with the inverse of the frequency of F1. Ladefoged at this point notices that *"the labels high-low and front-back should not be taken as descriptions of tongue positions. They are simply indicators of the way one vowel sounds relative to the another. The labels describe the relative auditory qualities, not the articulations"*, **[58]**.

Of course this kind of mapping has serious limitations as it takes into account only two formants. Ladefoged has rightly observed that *"the vowel chart does not show anything about the variations in the degree of lip rounding in the different vowels, nor does it indicate anything about vowel length".* The F3 formant can significantly contribute to the degree of lip rounding and rhotacization, **[58]** and a vowel mapping can also be built using all three formants (**Figure 3-2**). Obviously the displays used in Video Voice (**§2.3.2-E**), and the one used in HARP (**§2.3.2-F**) suffer from the limitations above. The vowel display can be useful for teaching vowels, but it can hardly work with consonants. The reason is that the quality of consonants cannot be displayed accurately with only F1, F2 and even F3 formants. As Rabiner and Schafer observe, **[82]**, *"in most consonants the first formant is either not observable or at a very low frequency. The frequency of the first formant will be at a minimum in most consonants in which there is an articulatory closure, but the frequency of the second formant varies considerably".*

B. Normalised formant mapping with a neural network and a vowel triangle

Recently there has been a similar approach to create a formant map and enhance it with a more attractive graphical interface and a video-game style **[1]**. The formant plotting as authors explain, **[31]**, is based on a Time-Delay ADAptive Linear Network (TD-ADALINE). This network receives as input 16-dim PARCOR vectors. The PARCOR vectors, obtained by applying the gradient adaptive linear predictive lattice algorithm on the raw speech signal, are further processed to estimate LPC parameters. These in their turn are used to evaluate the transfer function of the vocal tract. The first two formants are extracted, normalised, and plotted into a vowel triangle, its vertices being [-1, 1] (/i/) [-1,-1] (/u/) and [1,0] (/ɑ/) (**Figure 3-3**).

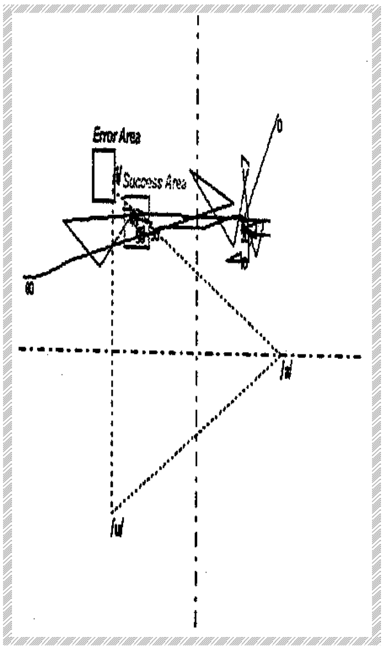


Figure 3-3 *The vowel triangle and the visual representation of the word ship. Scanned from* **[1]**.

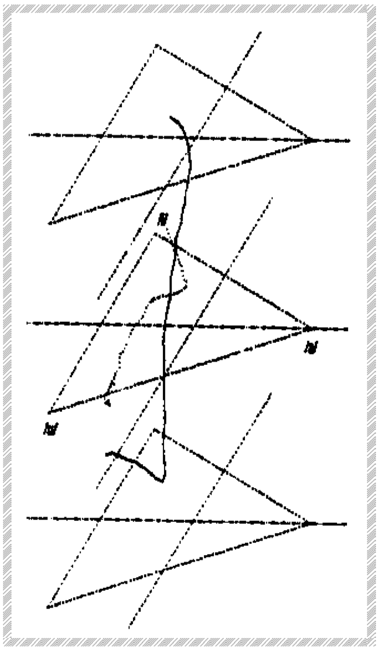


Figure 3-4 *3D mapping of the liquid and surrounding vowel sounds for the English word /yellow/ Scanned from* **[1]**.

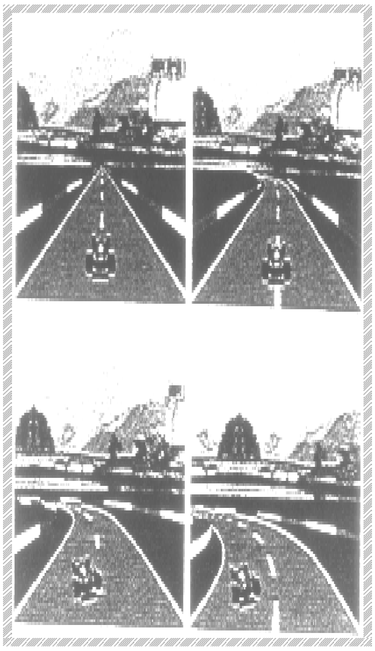


Figure 3-5 *A goal/target speech training display. Scanned from* **[1]**.

These x-y coordinates are used as the 2D output vector of the TD-ADALINE network. This mapping technique has been applied to generate x-y plots for different vowel-like sounds, mainly vowels, glides and diphthongs. A 3D mapping was also attempted by extracting the first three formants (**Figure 3-4**). The recently reported system, **[1]**, based on this technique also uses evaluative feedback in the form of a score that is translated into a visual representation of a goal/target to achieve (**Figure 3-5**). The authors claim that the technique *"is highly efficient in producing meaningful representations of different speech traces"*, and that it is intended for Computer-Aided Language Learning, **[1]**. Nevertheless we notice that the visual representation of the different speech traces is rather incomprehensible. Also the strong dependence of the system on F1, F2 and F3 formants makes it hard to apply it in non vowel-like phoneme categories.

C. Centroids frequency and standard deviation of spectral distribution

Wrench et al., **[100]**, take into account the number, frequency, and amplitude of both formants and anti-formants (poles and zeros). This technique uses centroid analysis to identify two peaks and troughs of spectrum (**Figure 3-7**). The two axes of the display correspond to the frequencies of the two spectral peaks, the horizontal with the lowest frequency and the vertical with the highest (**Figure 3-6**). Targets in the form of ellipses can be set by the therapist for the production of voiceless fricatives. The reference targets are based either on the patient's own speech target recorded pre-operatively or on an average of all patients' speech. The display has been applied for the assessment of segmental quality in voiceless fricative production of patients who have undergone intra-oral surgery. The visual feedback during speech production of the patient is instant and appears as coloured dots. Periods of silence or voiced speech produce nothing on the screen as well as fricatives with no spectral peaks.

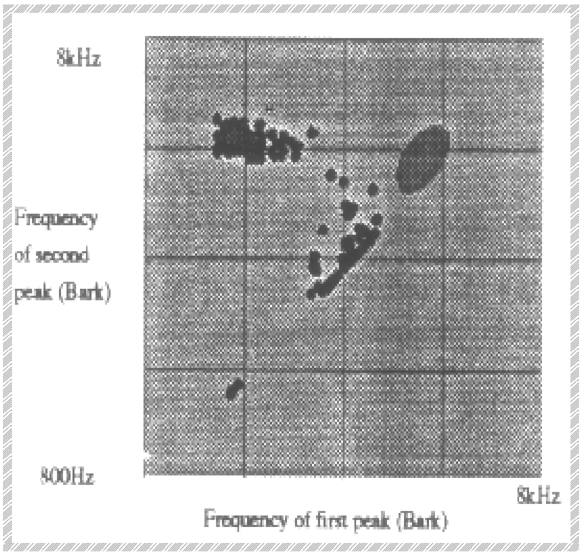


Figure 3-6 *Speech training display showing the target with the ellipse and the attempts of the client with the black dots. Scanned from* **[100]**.

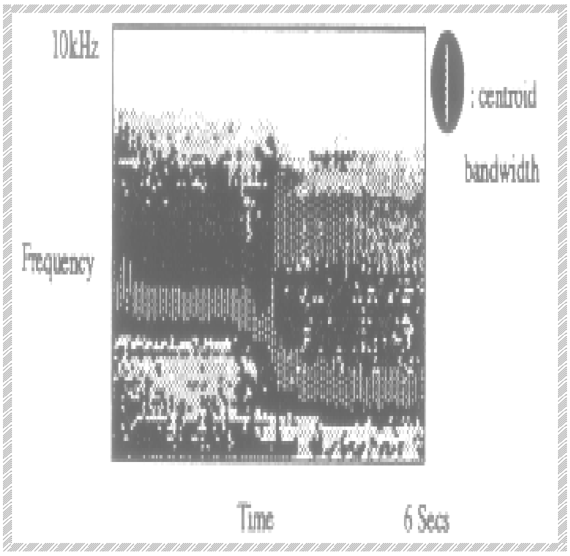


Figure 3-7 *Spectrogram showing dual centroid analysis of /s/. Scanned from* **[100]**.

3.1.2 Multivariate sound patterns visualisation

The second category of two-dimensional visual displays is based on multivariate data visualisation techniques. These techniques enable us in general to visualise the information contained in a high-dimensional data set such as cluster tendency, cluster shape, and inter-pattern similarities. Nevertheless, the relationship between the visual information provided and the goal of providing meaningful and instructional feedback to the user is not clear. Visualising the hidden structure of our input data space is not our primary objective, instead we want to provide visual feedback that is based on contrastive, consistent, and easy to interpret 2D visual patterns, requirements {C}, {D} and {E}. Methods which attempt to fit that description and which have been used in practice to implement 2D speech training displays can be categorised as follows:

Supervised non-self-organised techniques	**Unsupervised self-organised techniques**
Affine transformation (MSECT)	Principal Component Analysis (PCA)
non-linear linear transformation (NLLT)	Sammon neuro-mapping (SAMANN)
	Self-Organising Maps (SOM)

Table 3-1 *Taxonomy of multivariate data visualisation techniques in speech training*

We start by describing the unsupervised self-organised techniques as these were the first method of this kind applied to visualisation of speech on a 2D display.

I Unsupervised self-organised techniques

With the term unsupervised we denote that the category, in our case phoneme class information, for our data is not known a priori. With the term self-organising we want to make it clear that the layout of the phonetic map is created automatically and depends on the mapping technique.

A. Self Organising Map (SOM)

Although the term SOM is often used to refer to Kohonen's self-organising map **[56]**, the broad definition is that of a non-linear, unsupervised, projection from a d-dimensional input space to a 2D array, **[65]**. In SOM, we have a set of reference or 'codebook' vectors ($\mathbf{m}_i$) in the high dimensional data space, initially distributed randomly and associated with the nodes of a 2-D array. During learning, those nodes that are topographically close in the array, depending on the neighbourhood kernel ($\mathbf{h}_{ci}$)

shape and size, will become tuned and sensitised to input vectors (**x**) from the data space that are close to each other according to the Euclidean distance (|**x-m**|).

$$m_i(t+1) = m_i(t) + h_{ci}(t) \bullet [x(t) - m_i(t)] \quad \text{and} \quad h_{ci} = a(t) \bullet \exp\left(-\frac{\left\| r_c - r_i \right\|^2}{2\sigma^2(t)} \right)$$

where $c = \arg\min_i \{\|x - m_i\|\}$ and r_c , r_i are the radius vectors of nodes c and i respectively. After learning is finished, the distribution of the codebook vectors approximates the distribution of the input space. Clusters on the 2D array are self-organised and can be visualised by computing the distances between the codebook vectors of adjacent nodes, a technique known as 'Umatrix', **[49]**.

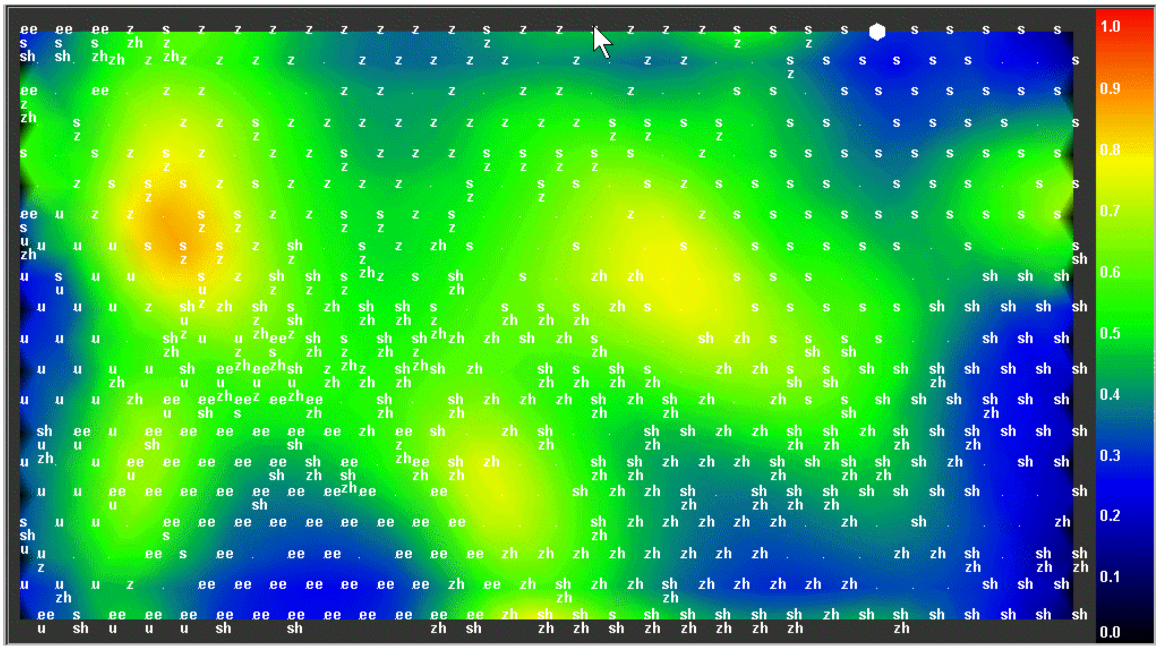


Figure 3-8 *Umatrix interpolated visualisation of SOM using the Neural Network Tool (Nenet v1.0) of Kohonen's team.*

Figure 3-8 shows a 20x30 SOM array with a multitude of colours that are interpolated between the medians of the calculated *Umatrix*. This SOM has been trained and tested with the data set of Table Appendices-20 that we used also for the 'iso2vo4frlvq' map of OLT. During the testing phase, a visual representation of an utterance can be obtained by plotting and connecting the successive positions of the nearest neighbour units activated by the sequence of the feature input vectors of the speech signal. The result is a speech trajectory like the one shown in **Figure 3-12**.

It was Kohonen who first speculated that this kind of display *"can be useful for speech training and therapy"*, **[56]**. Then Reynolds and Tarassenko implemented the 'Visual Ear', **[84]**, as an enhanced version of Kohonen's SOM. The problems they tried to solve were related to the smoothing of the trajectory (**Figure 3-9**) and the temporal variation due to the duration of an utterance (**Figure 3-10**). They also suggested a method for distinguishing visually the stationary and transient segments of an utterance. The 12x12 SOM they used included a set of phonemes, not more than four, of small words like 'way', 'were', 'able', 'mash'. The displays were used mainly for comparison of speech trajectories. The authors claimed that *"foreign language learners, or the post-lingually deafened can practise pronunciation by attempting to reproduce a variety of target trajectories which are derived from a large 'on-line' dictionary of speech sounds assembled by a normal speaker"*. To the author's knowledge, the *Visual Ear* project has not been continued since then and has not been evaluated or tested with real problems.

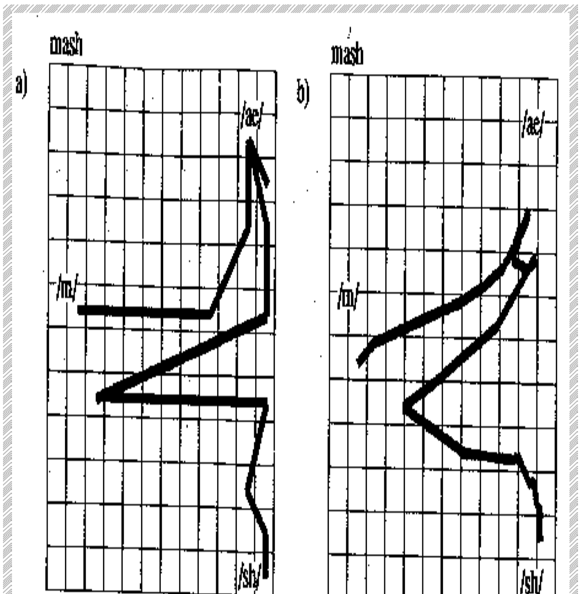


Figure 3-9 *Visual Ear trajectories for the word 'mash'. (a) Original, (b) Smoothed. Scanned from* **[84]**.

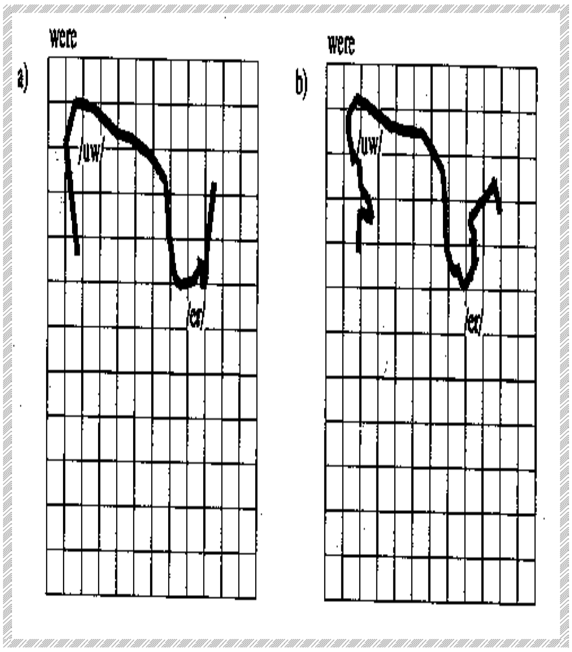


Figure 3-10 *Visual Ear trajectories for the word 'were' uttered (a) Quickly, (b) Slowly. Scanned from* **[84]**.

With the same technique Kohonen's team has explored the coarticulation phenomena present in the production of a vowel following a fricative, **[59]** and the difference between normal and dysphonic voices on the basis of the spectral composition, **[85]**. They showed that a SOM can successfully distinguish and visualise the different classes of sound.

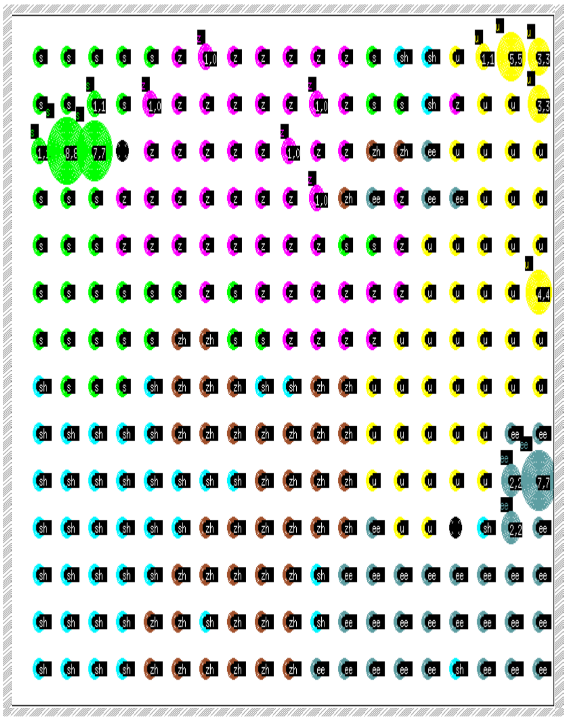


Figure 3-11 *SOM mapping for the utterance /ee-s-u/ of a normal speaker*

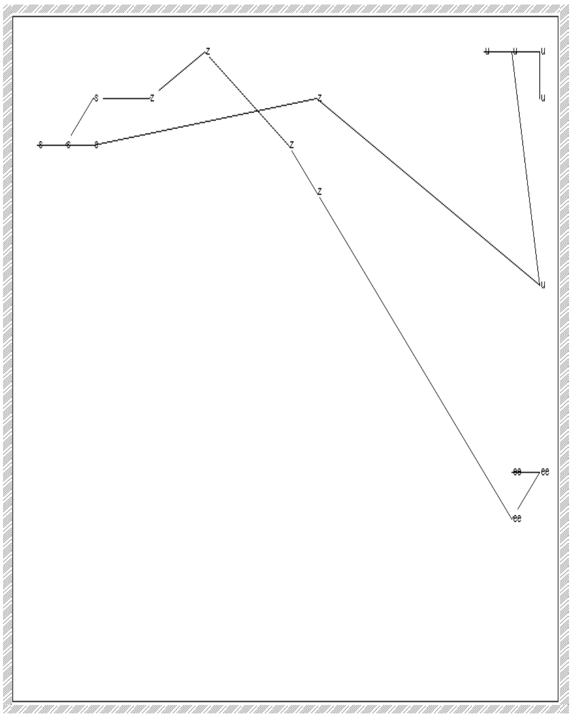


Figure 3-12 *SOM trajectory of the utterance /ee-s-u/ of a normal speaker*

We experimented a lot with SOM visualisation of the speech events at the beginning of our research. VAHISOM, **[36]**, an interactive visualisation and exploration tool of trajectories and SOM phonetic maps was used. We studied SOM with different dimensions, different training parameters, and different numbers of phonetic categories. These revealed problems related to the following important observations:

(a) The array of vectors used for the mapping is discrete. Essentially, the map averages over any series of input patterns by finding the closest codebook vectors. Bishop points out that *"previous attempts to consider magnification factors, the extent to which the area of a small patch of the latent space of a topographic mapping is magnified on projection to the data space, for the SOM were hindered because the manifold is only defined at discrete points (given by the reference vectors)"*, **[10]**.

(b) The confusion between the different phone classes increases at the borders of the class. In general, there is not a guarantee of well separable classes and the self-organisation of the layout can even result in more than one cluster with the same phonetic category. Even in the optimum case where all classes are well separable there will not be a clear visual interpretation or representation about the 2D borders of any sound cluster, or how it is separated and distinguished from all the other neighbour sound clusters on the phonetic map. The *Umatrix* visualisation can help greatly in that case.

(c) Increasing the dimensions of the map can result in empty spaces because of units that have not been tuned, or bigger confusion and scattering of the different phonetic classes represented on the map.

(d) Attempts to train SOM with parts from coarticulation between vowels and fricatives resulted in great confusion with the in-between non-transient phone classes area. This is also related to small utterance trajectories that are complicated and problematic in the segments of coarticulation, the transient parts from one phoneme to another. The self-organised layout in that case is not particularly logical or comprehensible, e.g. consider the layout of **Figure 3-11** and the trajectory in **Figure 3-12**. The units of the map representing coarticulated segments of speech for unvoiced sibilant fricatives are intermingled with those representing the classes for voiced fricatives.

B. Sammon neuro-mapping (SAMANN)

Sammon mapping, **[93]**, similarly to SOM, is *"a non-linear and unsupervised projection technique that attempts to maximally preserve all the interpattern distances*
If y(μ) $\in \Re^m$, *μ=1, 2, ..., n* patterns, *d(μ,v)* distance of μ, *v*, in the projected space (*m<d*) and ξ(μ)$\in \Re^d$, *μ=1, 2, ..., n* patterns, *d^*(μ,v)* distance of μ, *v*, in the input space (*d>m*).

$$E = \frac{1}{\sum_{\mu=1}^{n-1}\sum_{\nu=\mu+1}^{n} d^*(\mu,\nu)} \sum_{\mu=1}^{n-1}\sum_{\nu=\mu+1}^{n} \frac{\left[d^*(\mu,\nu)-d(\mu,\nu)\right]^2}{d^*(\mu,\nu)}$$

Sammon's stress E is a measure of how well the interpattern distances are preserved when the patterns are projected from a high dimensional space to a lower dimensional space. Sammon used a gradient descent algorithm to find a configuration of n patterns *in the m-dimensional space that minimises E.* Unlike SOM Sammon's algorithm *has no generalisation capability. To project new data, one has to run the program again on pooled data (old data and new data)"* **[65]**. For this reason a new on-line process called Sammon neuro-mapping (SAMANN), **[65]**, was developed using a feed-forward neural network to learn the non-linear relationship between high-dimensional data vectors and 2D points on the lattice. Nagayama et al., **[70]**, demonstrated this technique earlier by creating a phonetic map of the five Japanese vowels. There have not been any references for application of the display in real conditions of speech training.

Mainly because of the problems we referred to in the previous method, SAMANN was the second step in our research to develop a user-interactive interface to experiment with the visual feedback provided by a 2D phonetic map built with this technique. For that purpose, the OLT interface was built, **[37]**, the first version and model of our current OLT toolkit (OLTK). The learning vector quantisation (LVQ) method of Kohonen et al., **[57]**, was used to obtain a set of codebook vectors similar to SOM. This set of vectors was plotted on a scaled 2D lattice with Sammon mapping. Finally a back-propagation feed-forward artificial neural network (ANN) was trained to learn the non-linear relationship between the reference codebook vectors and the 2D points on the lattice. Alternatively, the original set of data vectors was used for training and the *LVQ* codebook set was used for labelling purposes.

One of the problems encountered in the programming of the user interface by that time was the dependence of our software on the *ANN* program for the real-time processing of speech input. Initially we used the 'Stuttgart Neural Network Simulator' (SNNS), **[106]**, and we had to convert the trained net into 'C' code, then merge the piece of code into our program, and finally recompile in order to built an executable that was highly depended on the type and structure of the *ANN* used. Training of the *ANN* with the SNNS interface was taking a long time to finish, and each time that we had a different speech data set or needed a different net we had to recompile and build again our program.

The map in **Figure 3-13**, was created with training data from four utterances of fixed context, eleven repetitions each, of the form /ee X u/ where X is /s/, /ʃ/, /z/, /ʒ/. The speakers recorded were four male English adults. From the SAMANN mapping of

Figure 3-13 and **Figure 3-14** we can observe several similarities and differences with the SOM mapping of **Figure 3-11**.

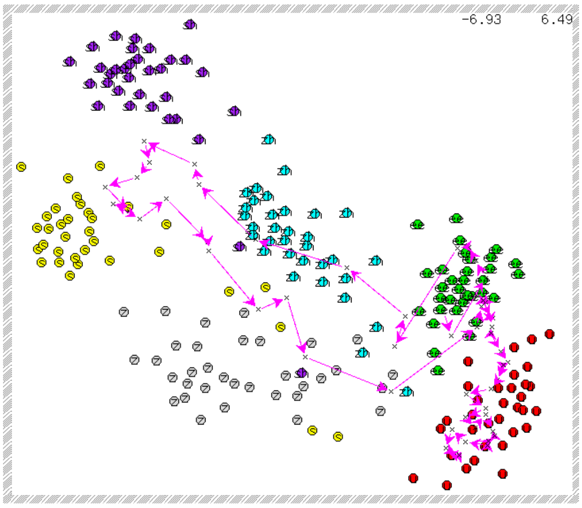


Figure 3-13 *SAMANN mapping for the utterance /ee-s-u/ of a normal speaker.*

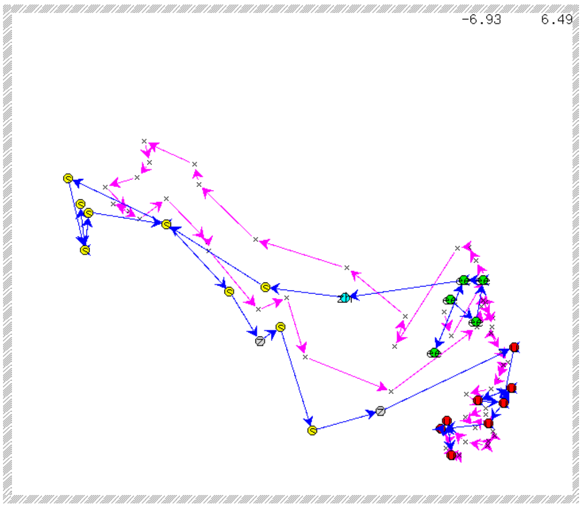
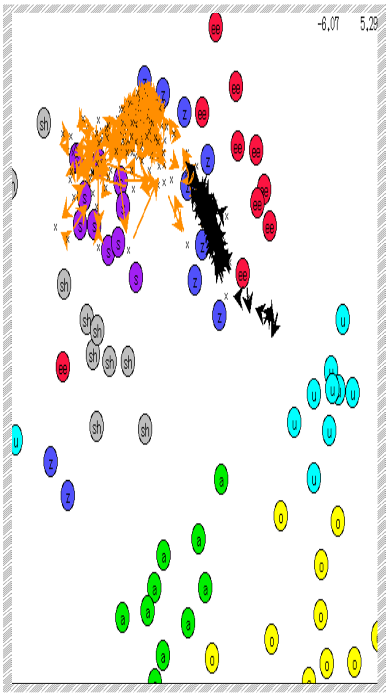
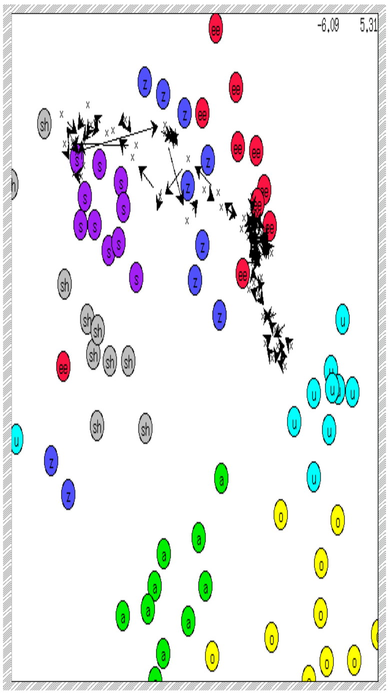


Figure 3-14 *SAMANN mapping vs. nearest neighbour mapping for the utterance /ee-s-u/ of a normal speaker*





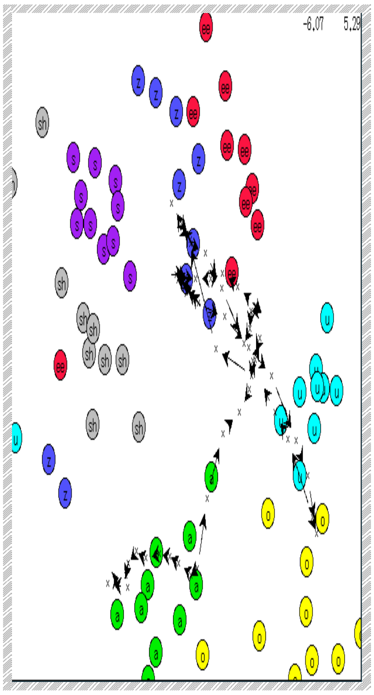


Figure 3-15 *Normal (black) vs. abnormal (orange) trajectory of the isolated sound of phoneme /z/*

Figure 3-16 *Abnormal speech trajectory of the utterance "a zoo" recorded from a child*

Figure 3-17 *Normal speech trajectory of the utterance "a zoo" recorded from a child*

First, the self-organised layout in both cases arranges the vowels close to the voiced sibilant fricatives, /z/, /ʒ/, and the unvoiced sibilant fricatives further away, preserving the property that close points correspond to phonetic similarity. But the lattice is continuous and can be defined at any point, not discrete as in SOM. The neural network learns to interpolate between the data vectors of each class. Nevertheless, concerning the visual clarity and interpretation of the overall image of the map as well as the plotting of trajectories, problems (b) and (d) mentioned with SOM still hold true. Later on a preliminary experiment was set on to test this old version of OLT with the misarticulated sibilant fricatives of a seven-years-old girl, **[38]**. The map was built from recordings of 9 female normal children, each repeating 5 times utterances of the form (/a/ – F – V) where F is /s, ʃ, z/ and V is /i, o, u/, e.g. "a sea, a sheep, a shoe, a zoo". Although the preliminary results reported with the child were promising, the therapy could not be carried on, because the child soon became disappointed and bored mainly because of the quality of the feedback provided, the appearance of the display and the limited motivation and reinforcement abilities of the interface. This is not surprising if we look at some of the results shown in **Figure 3-15**, **Figure 3-16**, and **Figure 3-17**.

C. Principal Component Analysis (PCA)

Another technique often used for multivariate data visualisation is the PCA. The PCA, like SOM and SAMANN, is unsupervised but performs a linear orthogonal projection. *"The 2D PCA map produced by the PCA network is spanned by two orthogonal vectors (in the original space) along which the data have the largest and the second largest variances*, and as such it can be considered similar to Sammon mapping. If $\boldsymbol{\xi}_{\boldsymbol{\kappa}} \in \Re^{d} = (\xi_{\kappa 1}, \xi_{\kappa 2}, \ldots, \xi_{\kappa d})^{T}$, and $\mathbf{O}_{\boldsymbol{\kappa}} \in \Re^{m} = (O_{\kappa 1}, O_{\kappa 2}, \ldots, O_{\kappa m})^{T}$, $\kappa = 1, 2, \ldots n$ and m= number of principal components, then from the linear transform $\mathbf{O}_{\boldsymbol{\kappa}} = \boldsymbol{\Phi}^{\mathbf{T}} \boldsymbol{\xi}_{\boldsymbol{\kappa}}$ we can obtain the *m* principal components; where **Φ** is a $d \times m$ matrix whose columns are *m* eigenvectors corresponding to the *m* largest eigenvalues of the covariance matrix *Σ"*, **[65].**

Ellis and Robinson attempted to construct a phonetic map such that, *"similar sounding phonemes will appear close together on a two-dimensional format, whereas dissimilar phonemes will be spaced far apart"*, **[19]**. They used the output vectors of a recurrent error propagation network phoneme recogniser and a measure of how the recognised phoneme is confused with the other phonemes. Then they performed PCA for all phoneme classes and plotted the results on a plane (**Figure 3-18**). As they notice, inspection reveals an area of confusion for most of the consonants plotted (**Figure 3-19**) and suggests that a non-linear transformation is required, **[19]**. Indeed, they used a self-organising technique to convert the PCA plane and make each phoneme occupy an equal area on a two dimensional grid arrangement (**Figure 3-20**), **[20]**. The grid was the basis of their phonetic tactile speech listening system, as an aid for the hearing-impaired. Recently a similar display was developed by Roy and Pentland, **[91]**, with the main idea that phonemes with high confusion were placed in close physical proximity. The estimated phoneme probabilities of the recurrent network are related to the brightness of each phoneme on the 2D display. No tests have been reported to assess the practicality of the above displays.

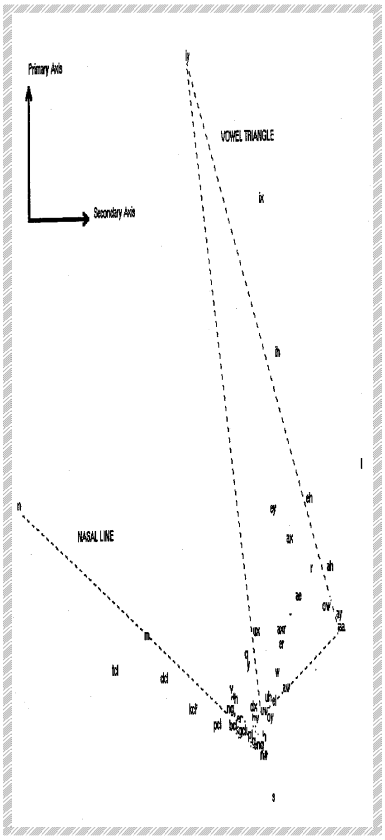


Figure 3-18 *PCA mapping of all the phonemes.*

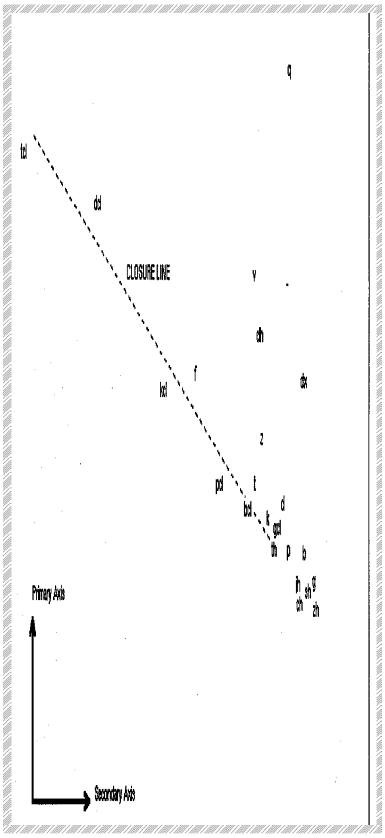


Figure 3-19 *PCA mapping and magnification for the area of the consonants.*

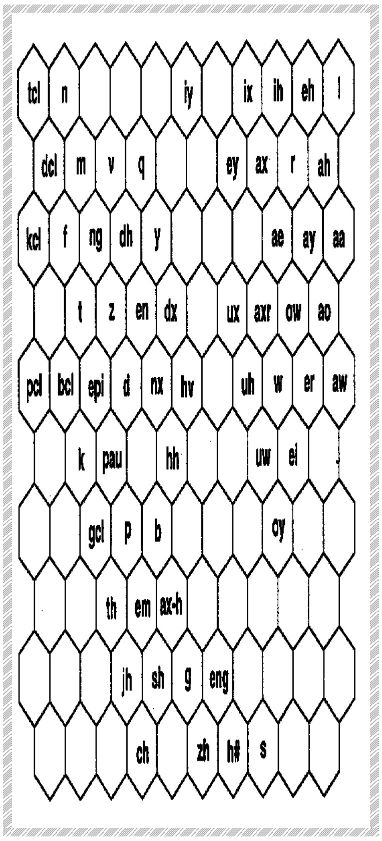


Figure 3-20 *Self-organising map for the English vowels and consonants.*

II Supervised non self-organised techniques

So far we have dealt with methods for multivariate data visualisation that are unsupervised. Nevertheless, in our case there is no reason, except for manual effort, to prevent us from labelling and constructing well defined sound classes. This is potentially advantageous both in terms of classification and visualisation as we can easily statistically model our data. As Mao and Jain notice, *"if the category information of the pattern is known, then it is more appropriate to use supervised learning"*, **[65]**. It is also a common property of all the previous techniques for the construction of phonetic maps that the input acoustic vectors that are close to each other according to Euclidean distance in input space, are mapped in two dimensional positions that are topologically close on the map. Nevertheless we have seen some of the problems associated with the self-organised layout. Fixing the centroids of the phone classes on the 2D lattice provides us with an interesting alternative technique. It is a natural property here, and a very desirable one, that the layout of the display can be designed by the teacher according to the needs of the client. The fixed phone classes can be laid from the teacher to suit a training problem -for instance front-to-back vowel order- in accordance with requirement {H}. It is also possible to build the layout in such a way that the transformed space correlates with a perceptual configuration of sound classes. The last point has been the motivation for the (MSECT) algorithm of Zahorian and Jagharghi.

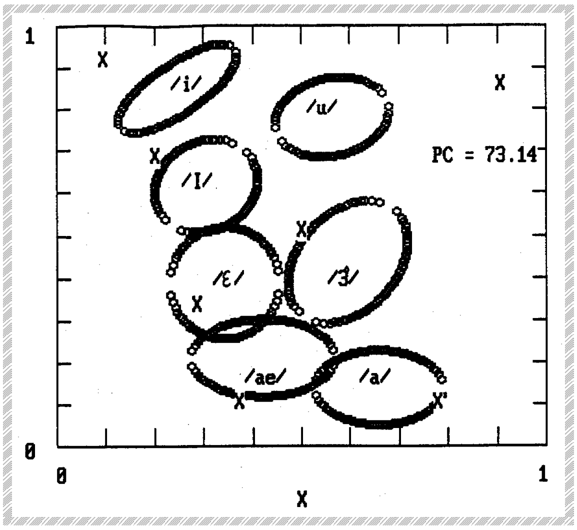


Figure 3-21 *Cluster plot from MSECT transformation.*

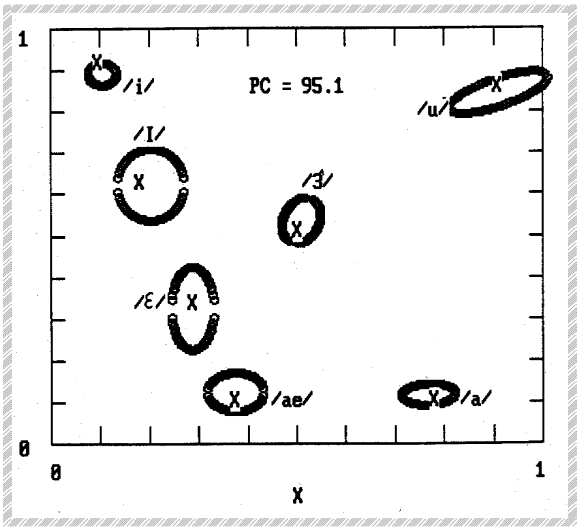


Figure 3-22 *Cluster plot from a neural network and MSECT combination*

A. Affine transformation (MSECT)

The term ‘affine’ refers to a general linear mapping of X onto X* described by the formula $X^* = XT + 1c'$ where T denotes an arbitrary linear transformation matrix and c' a *1 x n* row vector of constants. The mean-square error co-ordinate transformation (MSECT) computes the matrix T and the constants c' so as to minimise the total mean square error between specified target positions for each vowel and the actual transformed locations, **[104]**. The vowel targets shown in **Figure 3-21** were selected to correspond approximately to vowel centroids in a log F1 versus log F2 plot, **[105]**. As the authors report, this layout corresponds roughly to the front-back, low-high vowel dimensions.

A similar layout has been used in the ‘direct classification method’ of VSA speech training system (**§2.3.2-G**). The transformation is a linear combination of the parameters of the feature vector, using weights obtained from discriminant analyses and MSECT, **[79]**. In this way category centroids are computed in the 2D space, and the classifier works by calculating the minimum Euclidean distance of a transformed phoneme data vector from all phoneme cluster centroids (**Figure 3-23**, **Figure 3-24**).

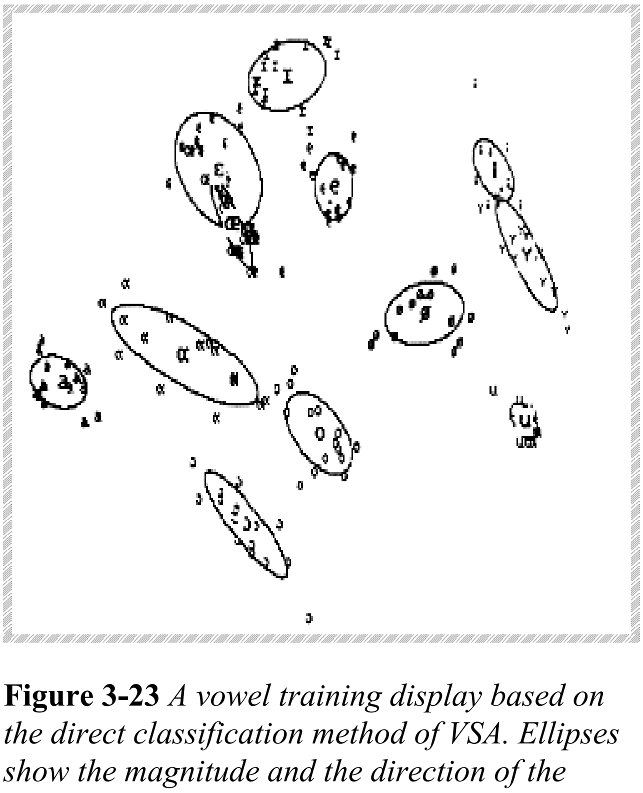


Figure 3-23 *A vowel training display based on the direct classification method of VSA. Ellipses show the magnitude and the direction of the variance within each phoneme cluster*

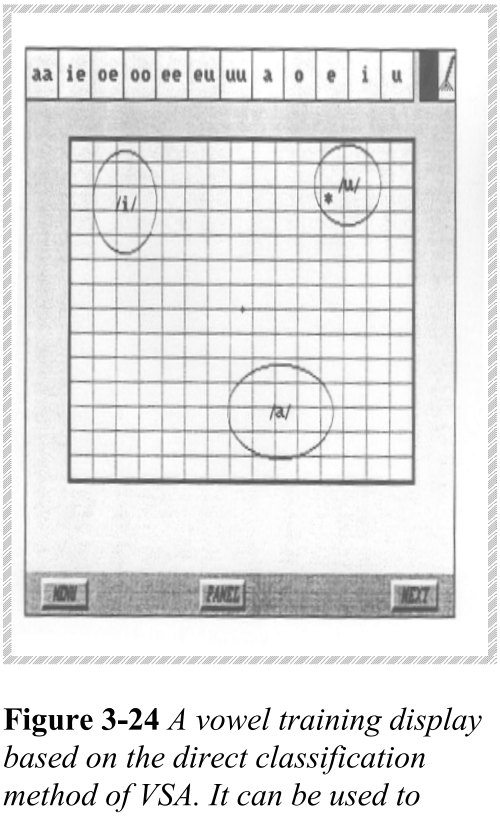


Figure 3-24 *A vowel training display based on the direct classification method of VSA. It can be used to explore the individual vowel space*

As Povel and Arends report, **[79]**, *"with the help of this display the deaf pupil can explore the vowel space and thus learn the relations between variations brought about in the vocal tract configuration and concurrent changes in the location of the point in the vowel plane. In this manner the child should be able to internalise the two dimensions of the vowel space that roughly correspond to degree of openness and back-front location of the tongue".* The therapist can gradually add more targets, thus enabling the child to form stable internal representations of the vowels.

B. Non-linear, linear transformation (NLLT)

In order to decrease the variance of the vowel clusters and better match the target position for each vowel Zahorian and Venkat, **[105]**, combined the MSECT transformation with a feed-forward back-propagation neural network trained as a classifier (**Figure 3-22**). The structure used is 12 discrete cosine transform coefficients as inputs, 25 hidden layer nodes and 10 outputs, one for each vowel. That network is used to drive the vowel bargraph display of their articulation aid, **[108]**. The height of the bar shows how correctly the vowel is pronounced and if more than one bars have non-zero heights it also shows the confusion with other vowels (**Figure 3-25**). For the ellipse display (**Figure 3-26**) an additional two-node layer is attached to the classifier output.

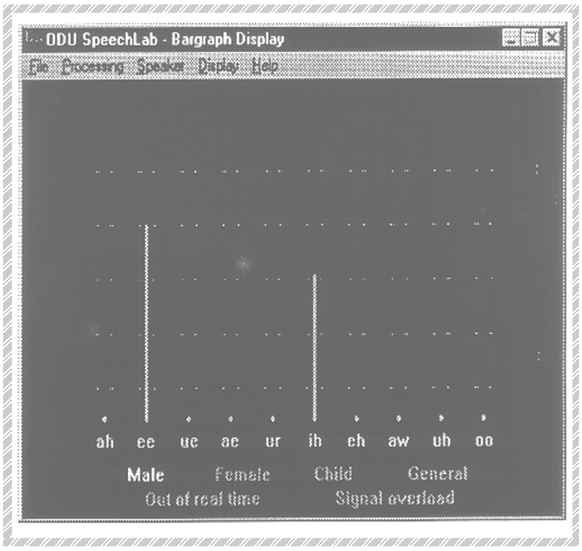


Figure 3-25 *The bar display of the vowel articulation training aid*

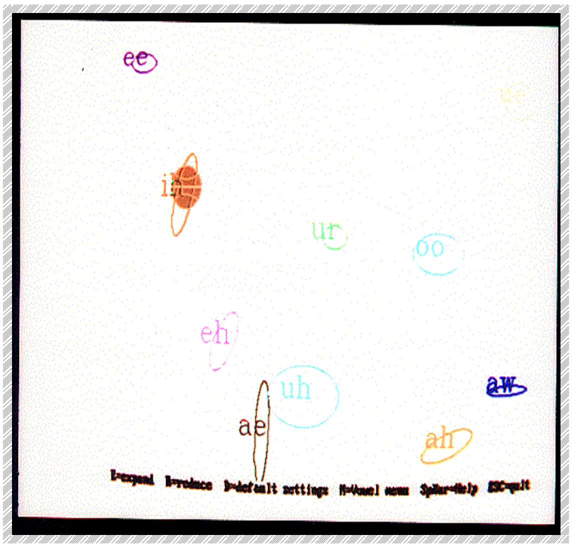


Figure 3-26 *The ellipse display of the vowel articulation training aid*

MSECT transforms the 10-dimensional neural network outputs to a 2-dimensional space with the predefined vowel target positions, **[104]**. According to the description of the display from Zimmer et al., **[108]**, *"in operation a correctly pronounced vowel guides a basketball icon into the ellipse region for that vowel and changes the ball's colour to match the ellipse colour. Incorrect vowel pronunciation causes the basketball icon to wander or appear in locations outside of the ellipses, or the ellipses of alternate vowels"*. A desired threshold can be defined as a Euclidean distance from the ellipse centre. Consequently, a counter can keep track of the number of utterances meeting or exceeding the threshold. There has been no significant formal testing and evaluation with hearing-impaired clients, although the authors report that both normal and hearing-impaired children improved their isolated vowel articulation, **[108]**.

3.2 Conclusions on two-dimensional speech training displays

From all the displays and methods we discussed in the previous section it is evident that accurate and simple 2D visual representation of the speech signal is not easy to accomplish. Although the method of multivariate visualisation of sound patterns seems advantageous in comparison with the formants mapping method, it appears to be problematic once an implementation of a speech training display is based on it.

3.2.1 Speech trajectories

In particular, the speech trajectories drawn on the display are not easily interpreted by the clients, thus violating requirement {C}. It is a difficult task to make a contrast between two different trajectories, as this requires smoothing and elimination of the temporal variation caused by many factors. Trajectories are actually dependent on the phonetic map drawn on the display which has to be as simple as possible, not cluttered, yet accurate and consistent in plotting sounds that have similar spectra on points with 2D positions that are close to each other according to Euclidean distance.

3.2.2 Visual representations of the sound classes

In order to simplify the problem and for maximum discriminative capacity we can select to map only a limited number of sound categories all of which we know a priori to be relative to the distribution of their class information. That means that a supervised technique should perform better for our purposes. We have also mentioned that it can be a desirable property if the instructor is allowed to select the layout of the sound classes. Thus, a non self-organised mapping should be preferred, where the option to specify the centroids of each cluster is given to the instructor. The mapping should also be continuous; and the space between two classes has to be defined. The sound classes have to be visually distinguishable and represented comprehensively as target areas for the client to aim at.

3.3 Important design aspects for 2D speech training displays

3.3.1 Target aiming

Aiming is in accordance with the principle of *visuomotor tracking* (**§1.6.1**). The stimulus consists of a static visual target that is present during a number of successive trials and has to be hit by single goal-directed movements. Articulatory targets can be visualised on a 2D computer display and represented as 2D cluster shapes, requiring a subject to hit the target, by producing an appropriate articulatory configuration. Therefore there should be a clear and comprehensible visual relationship of the articulatory gesture and the correlated phonetic area defined and displayed on the map, requirement {C}.

3.3.2 The consistency of mapping problem

Moreover this relationship, according to the requirements {C} and {D}, should be consistent and systematic. In practice, that means that small changes in articulation should result in plotting positions near the phone cluster area and larger changes should result in reaching 2D positions relatively far away from the phone target area. That brings into the discussion the requirement {F} about a metric to measure speech quality. Since visual contrast {E} is obtained from comparison with the norm of a model, we should measure how similar is the input acoustic vector to the sounds modelled. Nevertheless it is quite normal to expect productions which are highly deviant in comparison with those represented by the models of the 2D display. This raises the question of accuracy concerning the consistency for the mapping of the acoustic vectors onto two dimensions {H}. In other words the problem is how deviant the utterance of the speaker be from the target model of the training display, and what effect does this have on the mapping consistency. Since the models are dependent on the data set we use to train them, we called the problem 'Out Of Training Space' (OOTS) and it is discussed in details in **§4.6.4**.

3.3.3 Real-time processing

Another important design aspect for our speech training display according to requirement {A} is real-time processing. Practically this means, that there should be no delay between the client's response and the visual feedback presented on the 2D display. Despite the advances in microprocessor design (as we shall see in the next chapter) this still proves to be quite a computationally demanding task. Therefore the algorithms for the software implementation have to be as efficient as possible. For example, our first version of OLT training display, based on finding the winner unit and the associated 2D position of it by calculating the Euclidean distance for all the units of the SOM, showed that there are delays for processing speech input every 10msec. The real-time problem becomes more evident and harder to solve if, instead of plotting points, we have animation of a mobile graphics character, 'sprite'. In addition, one has to consider the computing time for each cycle of processing all other user interactions with the interface. All these issues are discussed in details in **§4.8.2-III**.

Chapter 4

*The Saviour of us all, the **Logos** of God, in His great love took to Himself a body and moved as Man among men, meeting their senses, so to speak, half way (3:15). When, then, the minds of men had fallen finally to the level of sensible things, the **Logos** submitted to appear in a body, in order that He, as Man, might centre their senses on Himself, and convince them through His human acts that He Himself is not man only but also God, the **Logos** and Wisdom of the true God (3:16).*

St. Athanasius-On the incarnation (Translation-C.S.Lewis)

OLTK - Models and algorithms

In this Chapter we discuss the software implementation of OLT, the Optical Logo-Therapy Toolkit (OLTK) application. The programming work was particularly intensive: more than one and a half years were devoted to the development of the new version of OLTK, not to mention the time devoted for accompanying programs. During this period the most important aspects of the software we faced were the real-time issues, the transformation from acoustics to points on the 2D display, and the out of training space problem (OOTS). Of course the whole program became rather complicated because it involved many other software design aspects such as the control of different drawing states, menus and options for the user, special graphics windows for results, and animation routines. The remaining of this chapter is organised around the "Speak and Look cycle" (**Figure 4-1**), the time sequence from digitally recorded sound to the visual feedback produced on computer display. We address the models and algorithms used in OLTK implementation of this cycle and finally we give some details of implementation itself.

4.1 OLT visual feedback

When we were discussing forms of sensory feedback other than hearing in speech training, we referred to the speak-hear learning cycle (**Figure 1-1**). Here we present an alternative route which we call the OLT speak and look *(*SpeakaLook) learning cycle. This is an outline for the software implementation of OLTK. The user speaks to a microphone and receives OLT visual feedback in real-time. Then s/he varies the articulation in order to produce systematic changes on the display and attain the target utterance. In **Figure 4-1** we can distinguish eight basic steps of that cycle which we are going to examine in detail in the following sections.

4.2 Speech data acquisition

The very first step in OLT *SpeakaLook* cycle is the digital signal processing of the sound card equipment during the recording or playback process of the software. The air pressure variations of the acoustic signal is transduced into electrical voltage differences from the microphone, and then an analogue to digital converter (ADC) stores the signal in a binary format. The reverse sequence occurs during playback.

4.2.1 Speech signal acquisition

Sometimes a quite neglected issue in speech training systems, but quite important, is the selection and use of the microphone. We followed the standards and the recommendation of the National centre for Voice and Speech (NCVS). This is a professional grade condenser microphone, omni-directional or cardioid, with a minimal sensitivity of -60dB. Two microphones were used both with a cardioid polar pattern. One hand-held, the PV-I Uni-directional microphone of Peavey company, and one mount-head Panasonic condenser microphone, WM-S10 with a pre-amp output connector. One problem we faced here is that when microphones are connected directly to sound cards in the computer, the signal picks up "noise" from the computer. This can affect the signal-to-noise ratio.

I Noise levels

We measured the noise level of the laptop computer sound card. This was done by disconnecting the microphone from the sound card and executing the recording process. Then we repeated the same procedure but by alternating the connection of two microphones. We obtained the following values for the peak amplitude where a level of 0dB is the absolute maximum of the peak amplitude of the signal before clipping occurs.

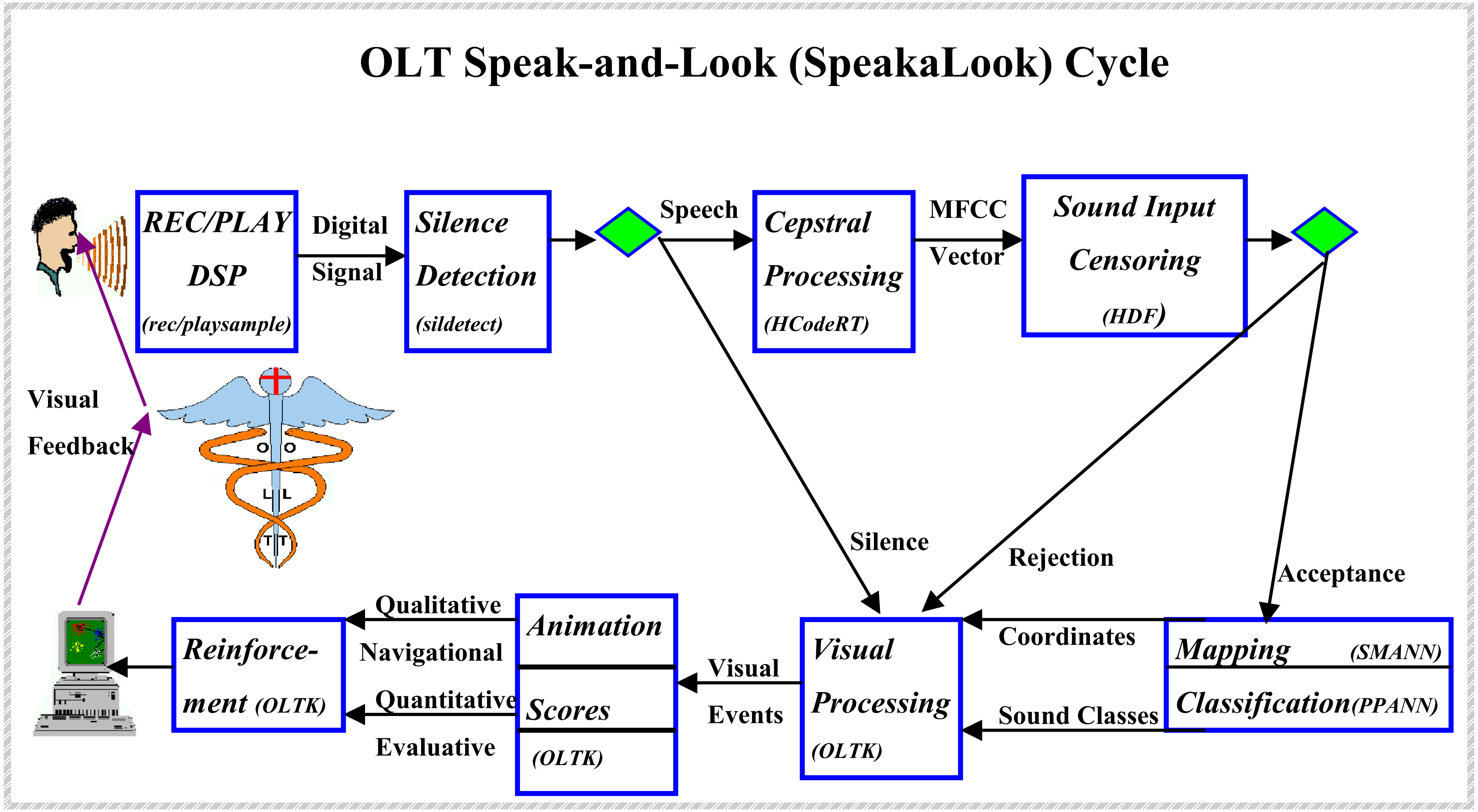


Figure 4-1 *The software engineering implementation of OLT speak-and-Look cycle.*

	Sound card alone	Sound card + microphone measuring silence
Minimum	-48 dB	-45 dB
Maximum	-33 dB	-27 dB
Average	-40 dB	-36 dB

Table 4-1 *Measurement of noise peak amplitude*

The measurement of silence has been done inside a computer lab where there was significant background noise coming from the fan operation of other computers. The average of the peak amplitude was only 4dB higher than the previous measurement. This is a good indication that our recordings made in a relatively quiet room have not been affected significantly by background noise other than that caused by the operation of computer hardware. The last has been the cause of distortion of speech signal due to frequent hard disk access (**Figure 4-2**); in that case the level of noise was increased periodically to -15 dB. Consequently the silence detection algorithm and evaluation are affected in that case. To prevent wrong evaluation from this effect we updated the scoring displays only if the duration of the non-silence signal exceeded a defined threshold, e.g. half a second.

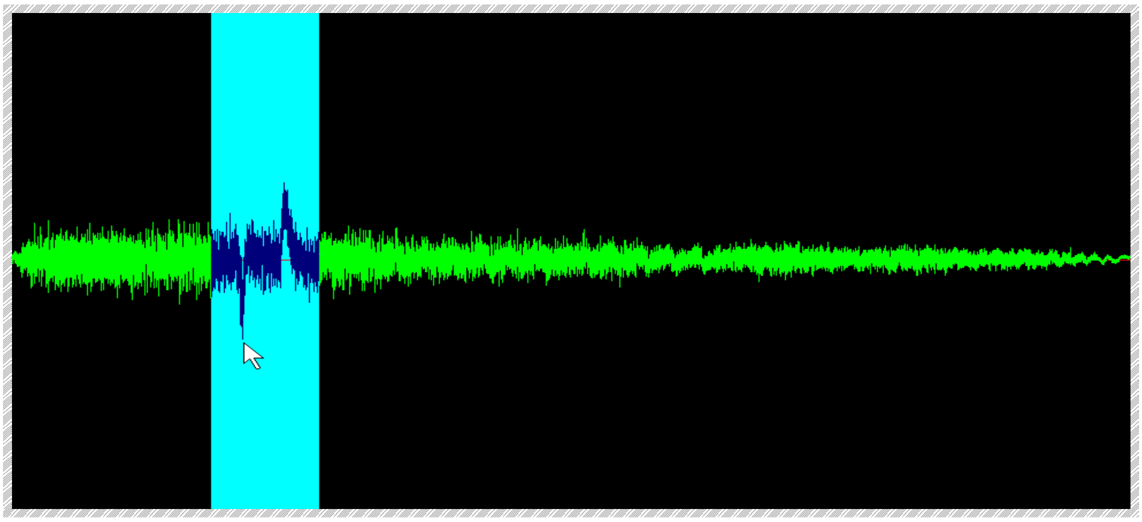


Figure 4-2 *Distortion of the speech signal of a fricative due to hard-disk access. Cursor and highlighted area show the distorted area.*

II Sampling

We chose to follow a sampling procedure according to the standards used in other similar speech commercial systems.The sampling rate was 16KHz and the resolution was 16bit. Samples were stored directly in the hard disk of the laptop computer.

4.3 Silence detection

The next stage in the *SpeakaLook* cycle is our silence detection program (*sildetect*, **§Appendix-III-A**). The algorithm is similar to the energy-based silence/speech detector of HTK, **[102]**. In a new environment where the recording conditions are different (background noise, computer hardware, sound card, microphone, recording volume, utterances recorded) the user can run *sildetect* and find what parameters are needed to tune silence detection to work more accurately. Obviously this is of critical importance for the consecutive processes on the cycle.

The example in **§Appendix-III-A** shows a run-time output of *sildetect*. First the program performs a recording of the background noise and calculates the logarithms of mean (silMean) and standard deviation (silStd) and the offset for a period of six seconds. This measurement period, 'FRAMESIZE', is equal to 100 frames of 60msec each. Then it prompts the user to make a loud utterance for the same period and calculates the log of maximum energy of the frame, 'logeng'. Execution proceeds with the *real_time_speech_detection* routine of the program. A number of samples are read either from a file or from *stdin* equal to *FRAMESIZE.* Then the frame's threshold is calculated, based on the following formula :

$$threshold = \frac{\log eng - silMean}{silStd}$$

According to the comparison of *threshold* with two levels, one for silence (*silThresh*) and one for speech (*spThresh*) which the user can set, the frame is identified as 'SPEECH', if *threshold* exceeds *spThresh*, 'SILENCE', if *threshold* is lower than *silThresh*, and 'UNDEFINED', if *threshold* lies between the two levels.

Finally, for speech detection we count the number of *SPEECH* and *UNDEFINED* frames and if they exceed a defined number of frames in sequence we write the speech frames to the output, *stdout* or a specified file. Otherwise we output for each frame the reserved value of the minimum short integer, 'MINSHORT'. This serves as an indication for the next process of the pipeline, the cepstral analysis, to propagate the information that silence is detected.

4.4 Cepstral analysis

In the source-filter model of speech production the vocal tract can be considered as a resonant system where the sound waves that pass through the vocal tract resonate well at some frequencies and not so well at others. The spectrum of the transmission characteristics of the vocal tract, the vocal tract transfer function, can be modelled by a filter with multiple resonators. The source contribution to the speech signal is made by the quasi periodic glottal air flow of the vocal cords that excite the vocal tract resonances.

The relative independence of the shape and therefore the resonances of the vocal tract from the excitation, means that these two components of the speech signal can be analysed separately from one another. Therefore, in capturing the time-varying spectral envelope for the speech it is desirable to reduce the effects of pitch in order to focus on the time-varying properties of the articulators. For such a purpose, we chose cepstral analysis to deconvolve the vocal tract filter and calculate the cepstral features. The features are orthogonal; and that simplifies pattern recognition, **[30]**. This has been also the preferred parameterisation method for most speech recognition systems. The algorithm we use, **[68]**, is implemented in the 'Hcode' program of HTK software, **[102]**. The basic stages of analysis are :

$$x(t) = e(t) \otimes v(t)$$

The speech signal in the time domain *(t)* can be represented as a convolution of the excitation source *e(t)*, and the vocal tract filter, *v(t)*. The first step is to divide speech data into overlapping blocks, windows, of 20msec, for every 10msec frame. Because of truncation errors caused by rectangular window function, we also apply a raised cosine, Hamming, window.

$$X(f) = E(f) \times V(f) \Rightarrow |X(f)| = |E(f) \times V(f)| = |E(f)| \times |V(f)|$$

The window of speech data is transformed in the frequency domain, *(f)*. The short-time spectra, *X(f)*, is computed by a 512-point FFT and the magnitude is taken.

$$|X(f_{mel})| = |E(f_{mel})| \times |V(f_{mel})| \Rightarrow \log|X(f_{mel})| = \log|E(f_{mel})| + \log|V(f_{mel})|$$

The next step is to bin the magnitude coefficients by correlating them with a mel-scale filterbank of 16 triangular filters and taking the logarithms. Mel-scale is based on pitch perception and is roughly linear below 1 kHz and roughly logarithmic above that frequency, **[30]**.

$$F^{-1}\left[\log\left|X(f_{mel})\right|\right] = F^{-1}\left[\log\left|E(f_{mel})\right|\right] + F^{-1}\left[\log\left|V(f_{mel})\right|\right]$$

The last step is to take the inverse discrete Fourier transform, IDFT. Since spectral values are real, the mel-scale frequency cepstral coefficients, (MFCC), are the cosine components of the IDFT. An additional step here is spectral smoothing; as Gold and Morgan notice, **[30]**, this is *"useful for reducing the effect of non-linguistic sources of variance in the speech signal"*. In practice this is accomplished by typically computing as many MFCCs as half of the number of filter bank outputs, **[103]** (8 in our case).

4.5 Sound input censoring

With the following models and algorithms we attack the *OOTS* problem (§**3.3.2**) and we define a suitable threshold for acceptance or rejection of the sound input, 'censoring'. Before we examine and comment on our method, it is essential to present briefly the steps for creating our data sets, since these are closely connected with the nature of the problem.

4.5.1 Speech samples acquisition

A good speech data set is always dependent on the quality of recordings. There are many conditions that make speech acquisition a difficult task, and in case we are to record words or small utterances we need to have control over several aspects of the acquisition. For example, consider disturbances in the signal, e.g. a cough, or mispronunciations. The researcher should be able to replay instantaneously any of the recorded utterances, and check whether the quality produced is good or not. If necessary the subject recorded should be instructed to repeat the utterance in question. Especially if the subjects are children and we record words, there should be a mechanism to indicate when recording starts and finishes. It is also desirable to show clearly, and for motivational purposes, what to record. Pictures of the recorded utterances seem to be a good approach. One should also consider how utterances should be stored. For our purpose, we wish them to be stored digitally on the hard disk under different files, with a unique identity tag composed of the name of the utterance, the name of the speaker, and a numeral if the utterance is repeated several times. Efficient handling of all the above aspects significantly decreases the time needed for preparing a good data set for further processing. The program we used to match all these requirements is the Speech Data Tool, (SpeDaTo, §**Appendix-III-C**).

4.5.2 Segmentation and Labelling

The next stage in building a speech data set, which is highly relevant with any supervised method of training for pattern classification and visualisation on two dimensions, is the segmentation and labelling of our data. Usually for speech recognition purposes segmentation and labelling is automatic, but for modelling articulation we preferred a higher precision manual method. Therefore we wrote a script (autolyre_segment, §**Appendix-III-D**), to speed up the whole process. This offers us also the advantage that we can double check the quality of speech recorded and discard certain segments or even whole utterances that will add 'noise' in our data sets (**§6.2.2III**).

4.5.3 Building the data sets

Once the utterances are recorded, segmented and labelled, the next step is to build the data sets. This requires first to combine the labelled segments of the utterance with the speech processing to produce labelled feature vectors. A demonstration of the execution of the program we built for such a task (ascii2dat) is in §**Appendix-III-E**, **Example 1**. This program has several user modes; a run-time output of the one we often used to build our data sets is in §**Appendix-III-E, Example 2**. In this mode, the program reads all the files specified in the masked argument option and sums all the labelled cepstral vectors for each sound category. Then it asks the user for an average number of vectors to build the training set, and another number to build the testing set from the remaining vectors. The choice of vectors is done in a random order. The whole process guarantees the creation of a balanced set of vectors, both in terms of equal number of representative vectors for each category and sufficient variance.

4.5.4 The speech data space

In **§3.2.2** we mentioned that we limit the number of sound categories to discriminate, and we try to collect speech samples for each class to have a natural cluster tendency and so that they can be well separable. For those reasons the data sets we built consist of a few vowels, 2-4, and a few sibilant fricatives, 2-4. A critical question to answer here is from what context to collect the speech samples. Two major possibilities exist: to take sustained phonemes, or to cut the segments from a context composed of syllables containing the phonemes. Clearly the second is much more difficult, as the variance of the samples increases due to coarticulation and assimilation at the points of transition.

To get an estimate of the distribution of the data in 9D space we consider the data sets, training and testing sets together, for the construction of two maps :

(a) from isolated phonemes, 'iso2vo4frlvq'
 (**Table Appendices-20, Figure Appendices-50, Figure Appendices-51)**

(b) from phonemes in context, 'freexu_lvq'
 (**Table Appendices-21, Figure Appendices-52, Figure Appendices-53**)

We plotted the data in two and three dimensions respectively (**Figure 4-3**). Notice that the two dimensional mapping of selected cepstral coefficients is similar to the Sammon mapping we have applied in the old version of OLT (§**3.1.2IB, Figure 3-13**).

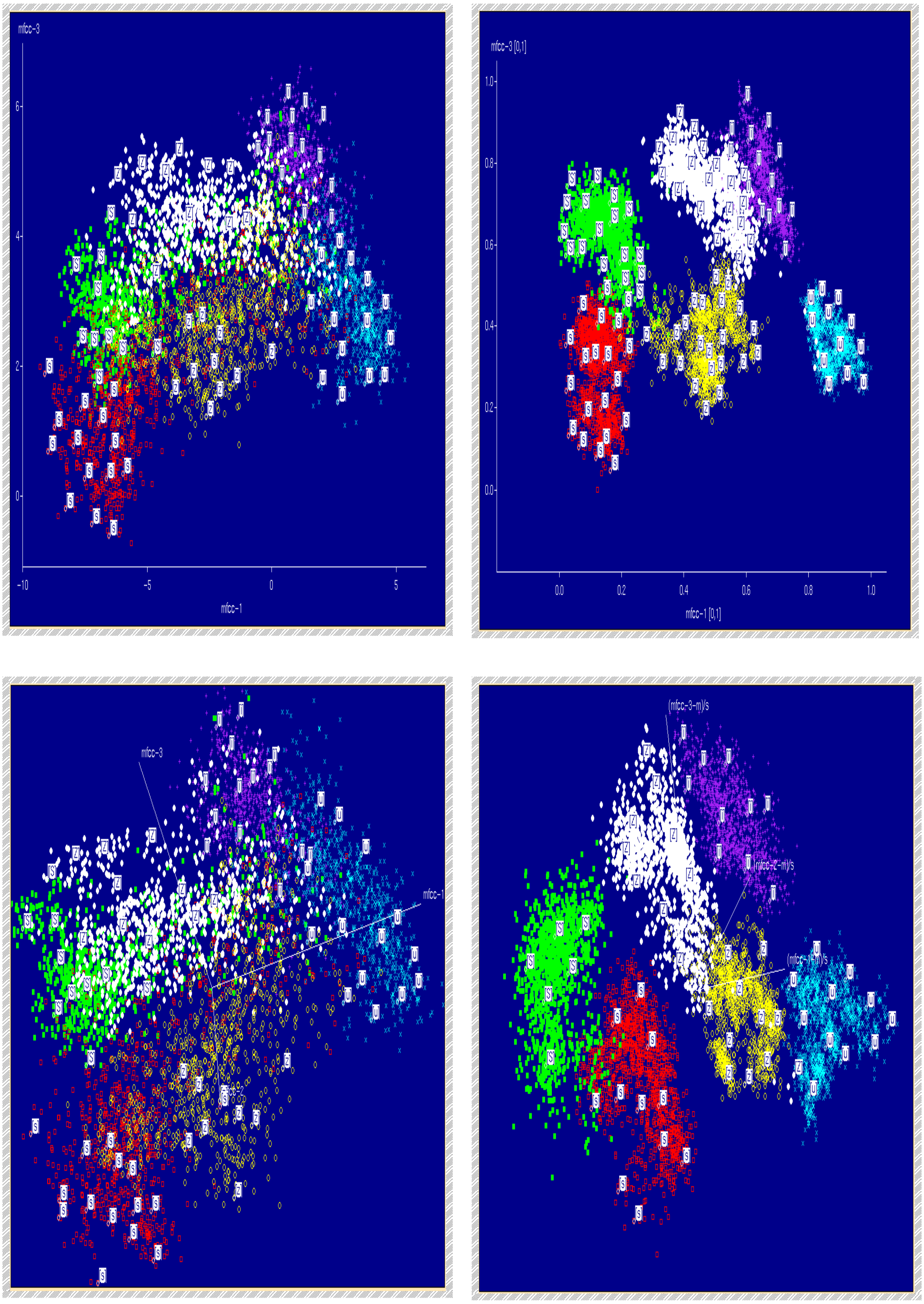


Figure 4-3 *Comparison of 6 phoneme classes, 4 sibilant fricatives (X), /s/, /S/, /z/, /Z/, and 2 vowels, /i/, /u/, between segments from /iXu/ context, (left half) and sustained production (right half) in 2D (upper half) and 3D (lower half). The axes represent the first two or three cepstral coefficients. Notice the significantly larger spreading of the fricative classes towards the vowel classes (left half).*

4.5.5 Modelling the sensitivity

I Modelling in speech training versus modelling in speech recognition

So far we have discussed many issues about how to create our data set and we have drawn attention to the within-class and between-class variability. The reason is that the data collected are going to be modelled statistically for comparison with the unknown incoming sound. This automatically raises the question of defining a metric. Before we talk about the metric it is important to highlight the difference from stochastic modelling in continuous speech recognition. Usually, despite the great variability of the corpus to train the recogniser, developers expect a specific accent and even highly probable states in the random sequence of the speech signal. Therefore the expected deviance of the recognised speech signal is assumed to fall inside a narrow band. Even if that does not occur from time to time, the lexical models can contribute significantly to recognising the spoken message. On the contrary, in speech training we expect highly deviant productions from the models. Anderson and Kewley-Port mention that *"the training of misarticulated speech calls for the evaluation of speech quality for speech that is non-standard with respect to productions of most speakers"*, **[2]**. The place and manner of the client's articulation is not known in advance, and even in the case where we have some information about the way the client articulates certain sounds, consistency is not guaranteed. So what is the key in modelling for speech training ?

II Defining a 'goodness-of-fit' metric for speech training

On one hand the model must allow for the allophonic variations, differences among the accents of normal speakers and for within-speaker variability. On the other hand the model must discriminate between those accents that substantially deviate from the target and those that fit into the range of acceptability, a 'screening' effect. As the developers of IBM Speech Viewer report, *"the key to the effective balance between tolerance of differences and discrimination of off-target performance is in the model setup process"*, **[48]**.

It is also important that evaluation of speech quality must be accurate {I} and agree with the judgement of the instructor. In a recent review of speech recognisers for speech training applications, recognition metrics based on nearest neighbour classification and spectra comparison were found to be very comparable to the performance of trained human ratters, **[2]**. Same authors report that a stochastic metric, like HMM confidence score, *"is not a 'goodness-of-fit' measure, so it is possible to have a high confidence score and yet a relatively poor match between utterance and underlying Markov model"*. Well known speech training systems that use spectra comparison and 'goodness-of-fit' metrics for evaluation of speech are used in ISTRA, and IBM Speech Viewer.

A. KNN based metric

We performed experiments with the old version of OLTK to test in real-time the nearest-neighbour rule with the construction of an *LVQ* codebook set, and Euclidean distances, **[37]**, **[38]**. Despite the processing power, the computational complexity of the algorithm was prohibitive, verifying the reports of other researchers, **[18], [30], [10]**. Even if we generalise to a k-nearest neighbour (KNN) rule, *"where the class of the new vector is decided by a vote of the class of the k-nearest neighbours, and the voting weight is determined from the distance of the neighbour"*, **[30]**, *"the resulting estimate is not a true probability density since its integral over all x-space diverges"*, **[10]**. In that case, this rule achieves, *"at most twice the error of the optimum classifier"*, **[18]**.

B. Relationship of KNN and Discriminant functions

Those reasons led us to think of a similar approach to compute our metric and satisfy the *goodness-of-fit* requirement. As such, assuming that the form of the underlying parametric density function for each phoneme class is a unimodal Gaussian, we can base our metric on the maximum hyperquadratic discriminant function (HDF).

The general multivariate normal conditional probability density function (pdf) in **d** dimensions for the class $\boldsymbol{\omega}_\mathbf{i}$ with covariance matrix $\boldsymbol{\Sigma}_\mathbf{i}$ and mean $\boldsymbol{\mu}_\mathbf{i}$ can be expressed as

$$p(\mathbf{x} \mid \omega_i) = \frac{1}{(\sqrt{2\pi})^d |\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \mu_i)^{\mathrm{T}} \Sigma_i^{-1} (\mathbf{x} - \mu_i) \right] \qquad (1)$$

$$r^2 = (\mathbf{x} - \mu_i)^{\mathrm{T}} \Sigma_i^{-1} (\mathbf{x} - \mu_i) \qquad (2)$$

$p(\mathbf{x} \mid \omega_i)$ is also referred to as the likelihood function. "*Intuitively we can think of a likelihood function as being a kind of a 'closeness' measure (if a particularly class-*

dependent density is closer to the new observation than the other densities, it will tend to have a higher likelihood)", **[30]**.

The term r^2 is called the squared Mahalanobis distance from **x** to **μ_i**

If we choose to use r^2 as a minimum distance classifier, manipulating the algebra, we get : $\arg\max_i (r_i''^2 = x^T \mu_i)$ In this case the minimum distance classifier is equivalent to a classifier based on the maximum dot product with the normalised mean prototype. In the general case of *KNN* rule to find the maximum dot product between a new vector and the stored prototypes can be viewed as *"determining weights for a linear recombination of the input features that will be maximum for the correct class"*. In statistics this is called a linear discriminant function g$_i$(x), and generates a hyperplane decision surface.

C. Metric based on hyperquadratic discriminant function

If we consider the Bayes rule for Maximum a Posteriori (MAP) classification based on the likelihood and the relative magnitudes of a discriminant function, we have

$$\arg\max_i \left(g_i(\mathbf{x})\right) = \arg\max_i \left(p(\omega_i \mid \mathbf{x})\right) \overset{Bayes}{=} \arg\max_i \left(\ln p(\mathbf{x} \mid \omega_i) + \ln P(\omega_i)\right) \overset{(1)}{\underset{(2)}{=}}$$

$$\arg\max_i \left(-\frac{1}{2} r^2 - \frac{1}{2} \ln|\Sigma_i| + \ln P(\omega_i) \right) = \arg\max_i HDF_i$$

The geometric interpretation of the formula above is that in the general case *"the decision surfaces are hyperquadratics, and can be any of – hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, and hyperhyperboloids of one or two sheets"*, **[18]**. Therefore the *HDF* function can serve both as an optimal classification rule and as a *goodness-of-fit* metric based on the maximum likelihood, if the statistical assumption for the unimodal Gaussian form of *pdf* for each class is justified. The prior class probabilities $P(\omega_i)$ depend on the fraction of samples from different phoneme classes in the speech data set.

D. Algorithmic and computational implementation

We implemented the *HDF* with routines taken from a pattern recognition toolbox called 'tooldiag', **[83]**. In this way we built a statistics library compatible with C++ compiler to serve our programs. This library includes, I/O functions, memory allocation functions and data types, and multivariate statistics calculations (means, covariances, confusion matrixes, and *HDF*).

4.5.6 Experimental results for HDF

To test the accuracy of the *HDF* metric and compare it with other methods we built two special modes in the *ascii2dat* program. The 'CLA' mode compares classification results between the *HDF* metric, an artificial neural-network (ANN), maximum a posteriori estimator (MAP), and *LVQ* method. The 'THR' mode tests the discriminating ability of the *HDF* to accept those input vectors that are sufficiently close to the feature vectors of the training space (valid samples) and reject those that do not 'fit' in the modelled phoneme classes of the training set (invalid samples).

I Classification comparison

We run the *ascii2dat* in the *CLA* mode (§**Appendix-III-E, Example 3**) to collect the classification results for the two different data sets shown in **Figure 4-3**. A summary of the results presented in **Table Appendices-2** and **Table Appendices-3** can be seen in **Table 4-2**. The optimum classifier was the *ANN* method in both sustained phonemes and phonemes in context; the next best was a combination of *KNN* method and *LVQ* template vectors. The *HDF* method came last, but the scoring differences from the other methods in sustained phonemes were insignificant, and only a 2%-4% decrease appeared in context classification.

Method	**Learning from**	**Sustained**	**In context**
ANN - 9x16x6	*Raw Training set*	*99.96%*	*94.23%*
KNN	*LVQ - 6x100 set*	*99.91%*	*93.95%*
HDF	*Raw Training set*	*99.90%*	*92.18%*
HDF	*LVQ - 6x100 set*	*99.86%*	*90.47%*

Table 4-2 *Comparison of the four classification methods against the two different data sets.*

As was expected, classification in sustained phonemes was much more accurate than in context: the rows of fricatives in **Table Appendices-2** and **Table Appendices-3** show a 6%-10% drop below the average, particularly in the unvoiced sibilant fricatives /ʃ/ and /s/. Indeed, if we have a close look at the visualisation of *freeXu* data set in **Figure 4-3**, and the mapping of the corresponding codebook set in **Figure Appendices-52**, we can see how much more the fricatives vary comparing to the vowels. This makes us more firm in the choice of selecting the better clustered sustained phonemes data to build our map.

The purpose of those results was primarily to examine the discriminative power of *HDF* for data that are expected to be close to the Gaussian densities of the phonemes. The conclusion is that for well-clustered data this assumption does not harm the classification accuracy. If the performance was poor, there should not be any reason to study further how *HDF* behaves with input vectors that are far away from the distribution.

II Acceptance/Rejection Threshold

Among other reasons we mentioned, the *HDF* metric is useful because we can set a threshold to either accept those input vectors sufficiently close to our mode, or reject them. Obviously the question here to answer is what it means in practice for input vectors to be sufficiently close to the models distribution, and what is the optimum threshold setting. If we consider the phoneme classes represented in our data set, we can say that the input of any other phoneme class should be rejected. Of course this is the easy case, because we expect these other phonemes to differ substantially from the phoneme models. Nevertheless in speech training this is a very naïve approach because of two factors: first, we expect an abnormal deviant accent that acoustically can be quite similar with a normal accent; and second, the models have to be broad enough to capture any variant accent that can be considered as normal production. We shall discuss this important topic again in §**4.6.4**, and we examine its effect in a case study and in the approach we followed in §**6.2.4-IV**.

Now let us solve here the easy part, thus discriminating between input samples generated by phonemes similar to those of the phoneme models classes and other samples generated by phonemes not seen in the training test. Two types of error can appear at any threshold setting of our metric method. Type I error measures the percentage of *valid* samples rejected, and type II error measures the percentage of *invalid* samples accepted. In other words, if the threshold is set in a very high value, the discriminant surfaces may capture only part of the models distribution or may not be broad enough to accept a normal deviant accent from the phoneme models. This results in rejecting *valid* samples. In type II error, the threshold is set in a very low value, which means that the hyper-space covered by the hyperquadratic surfaces is much larger than that of the distribution and can accept sounds that are too different from those our models consisted of: thus at that level it starts accepting samples considered *invalid*.

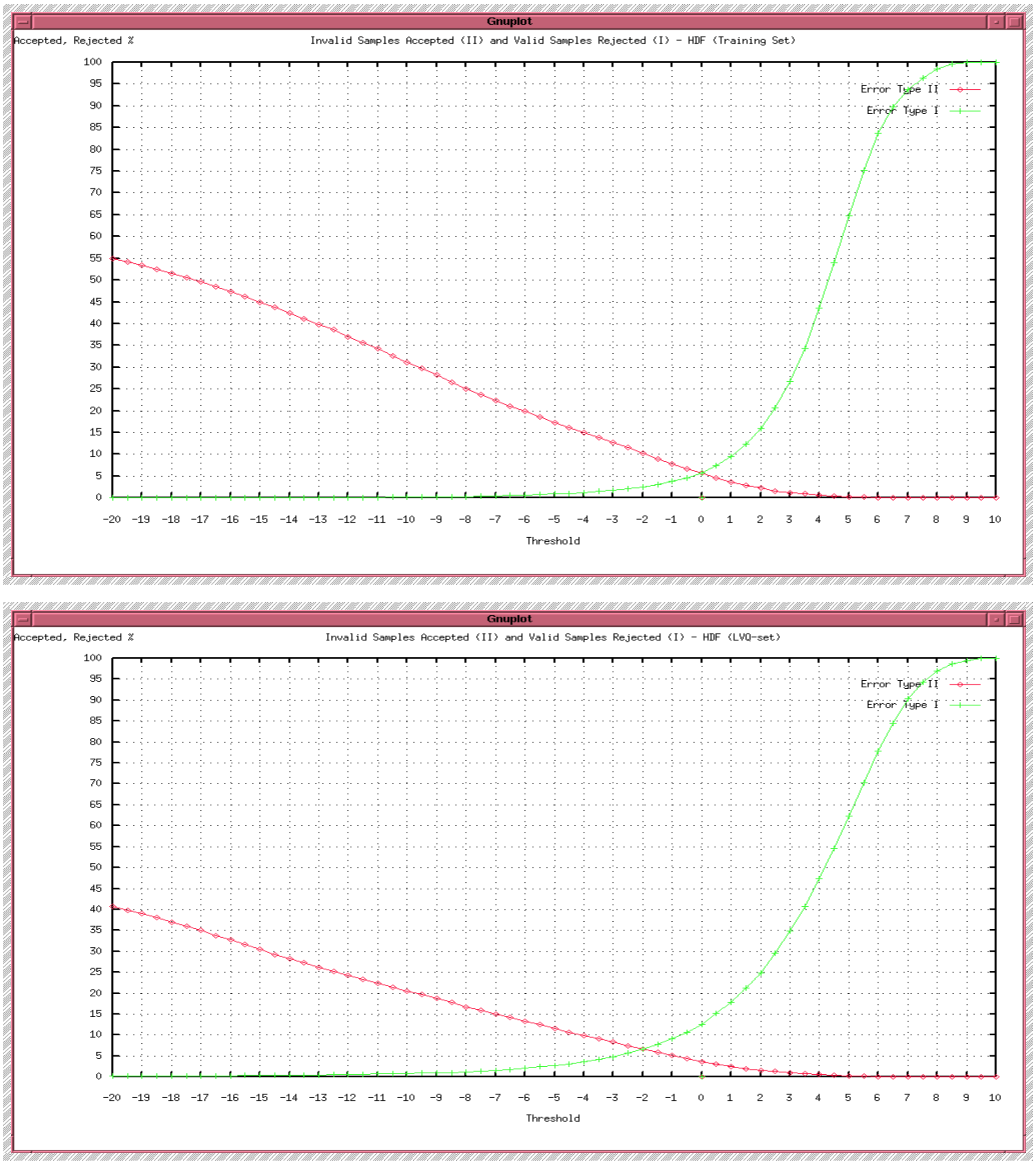


Figure 4-4 *Optimum threshold setting is at the crossing point with a minimum of 5% error for each of the error types for an HDF metric and raw data training set (upper graph) or LVQ (6x100) codebook set (lower graph).*

The optimum threshold level can be set once we know what kind of samples we consider invalid. For this purpose we labelled and analysed samples recorded from 8 normal English, adult, male, speakers pronouncing 9 sustained phonemes. We compared them with the 6 sustained phonemes of the *iso2vo4frlvq* map already used in the previous testing. For modelling of the *HDF*, both an *LVQ* and a non-*LVQ* set was used. The program to run the tests was again *ascii2dat* equipped with the *THR* mode

(**§Appendix-III-E, Example 4**). As we can see in both of the graphs of **Figure 4-4**, the error type I increases as the threshold is set at higher values, while the error type II decreases. The opposite happens for lower values of threshold. The optimum setting achieves a minimum error of roughly 5% for both types of error, and for both an *LVQ* and a raw training set. Nevertheless, we prefer the *HDF* metric with the *LVQ* set than *HDF* with a raw training set because, for lower values of threshold, error type II is less (15 degrees decrease of the threshold from the optimum position gives a 10% less error).

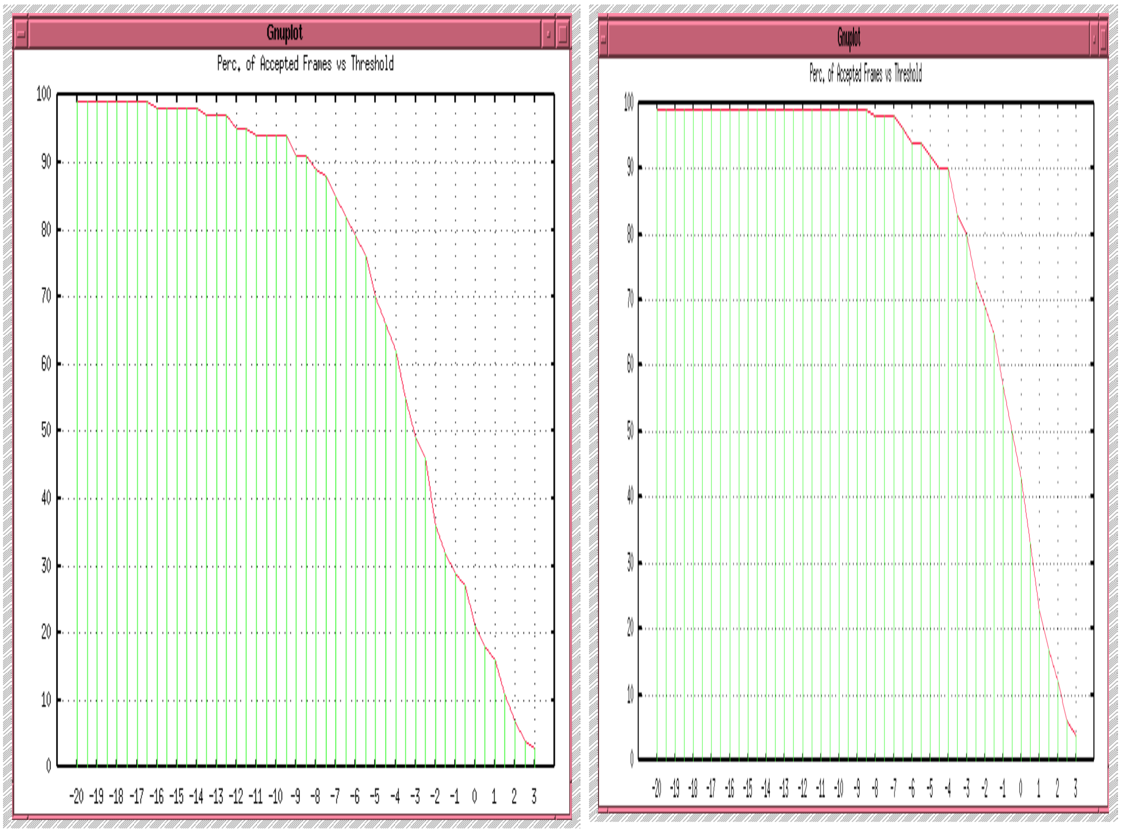


Figure 4-5 *Acceptance comparison with variable threshold for an abnormal /s/ sound. The left panel is the HDF with the LVQ set and the right panel is HDF with the raw training set.*

An additional argument concerns the smoothness of *HDF* metric. In **Figure 4-5** we can see that 80% acceptance for the samples of an abnormal /s/ frication is gained with the threshold at (–6) if an *LVQ* set is used. While the same sound with the same acceptance is gained with the threshold at (-3) if a raw data set is used. The metric in the first case varies more smoothly than the second.

In conclusion, the optimum discrimination of speech input presented to *OLTK* based on *HDF* metric and an *LVQ* set can be achieved at a certain level of the threshold, with a minimum error of 5%. This performance occurs only if we know ahead what is the expected data input in our system, both in terms of *valid* and *invalid* samples, and if there is sufficient contrast between the two. Any shift from this level of threshold means that the system tends to become either insensitive or oversensitive to the speech of the user. However, in practice we do not know what the speech input sounds like; and the only certainty is that we expect, depending on varying the level of the sensitivity, that the deviant accent will be either too easily accepted, or too often rejected. The first involves the danger that the visual feedback provided may be inconsistent and the evaluation of *OLTK* may be wrong; while the second can cause the client to become discouraged. Of course, especially with children we have to be cautious and avoid the discouragement. Moreover, there is also the likelihood that the sensitivity will fail to distinguish between normal and abnormal productions if the two are acoustically very similar (§**6.2.4-IV**). Actually these are two sides of the same coin, the *OOTS* problem that we examine and explain in detail in §**4.6.4**.

4.6 Mapping and Classification

So far in the previous sections we have discussed the usefulness of defining a metric and a threshold to control the sensitivity of the censoring for the accurate evaluation of speech quality. However this is only half of the story. The other half is related to the transformation from the 9D cepstral space to the 2D space of the phonetic map. In section §**3.3.2** we briefly mentioned that the influence of the sensitivity control on the consistency of the mapping is profound. Here we present all the steps for training the neural networks responsible for the mapping and classification and review the *OOTS* problem.

4.6.1 Designing the training and testing set

There are several issues one has to consider in designing the training and testing set to apply to the learning process of the *ANN*.

A. Training set representation

We have already noticed the importance of building our speech data space (§**4.5.4**) and modelling the sound categories (§**4.5.6-II**). The performance of *ANN* is highly dependent on these factors: after all the network learns according to the patterns which present to it. The above can be summarised by saying that every sound class must be represented adequately in the training and testing sets. This means that there should be enough statistical variation in our speech data.

B. Balancing the classes

Some consideration should be given to the number of patterns representing each sound class. Since all sound classes represented in our map are equally important, and since the client is required to produce a specific sound, the sets must be balanced. As Masters notices, *"when a network learns by minimising the mean error across the entire training set, the proportional representation in the training set can have a profound influence on the network's performance"*, **[66]**. Alternatively we could have used unbalanced sets; but then the learning algorithm has to be modified accordingly to include the prior probabilities (priors). In our software implementation of the *ANN*, *priors* are not explicitly set; therefore we equalise the number of vectors for each sound category.

4.6.2 Practical considerations in training ANN

We chose to experiment with the most wide-spread and well-behaved type of *ANN*, the feedforward, back-propagation, multi-linear perceptrons. *"The aim is to train the net to achieve a balance between the ability to respond correctly to the input patterns that are used for training (memorisation) and the ability to give reasonable (good) responses to input that is similar, but not identical, to that used in training (generalisation)."*, **[24]**.

The stages involved in training are three: feedforward of the input training pattern; backpropagation of the accumulated error; and update of the weights. The algorithm is based on the gradient descent technique known as steepest descent : $\Delta W^{(\tau)} = -n\nabla E\mid_{W^{(\tau)}}$

Where *n*, the learning rate, is the amount the weight vectors *W* are updated in the steepest direction in which the error *E* decreases towards the local minimum. The *(τ)* is called 'epoch'.

In both types of *ANN* we use, the factors that affect training are :

- **the number of samples in the training and testing set.**

According to Masters, **[66]**, *"the only way to prevent the network from learning unique characteristics of the training set, to the detriment of learning universal characteristics, is to flood it with so many examples that it cannot possibly learn all of their idiosyncrasies."*. A rule of thumb to compute the minimum number of training samples required is : Patterns = Weights/Error , **[24]**.

For example for a 9 inputs, 16 hidden units and 2 outputs *ANN* we have :

$(9+1)\times 16+(16+1)\times 2=194$ weights. If we assume a 0.01 error for the testing set, the minimum number of patterns will be 1940.

❖ **the learning parameters :**

♦ **number of hidden units.**

The training set size and the hidden-layer size are closely tied together. The number of hidden units must be such that *ANN* is powerful enough to learn the problem without the insignificant and irrelevant to the general population aspects of the training set. Therefore we prefer the smallest possible number of hidden units that gives us the minimum error in the testing set.

♦ **learning rate.**

If the learning rate is too small, convergence will be excessively slow, and we may be trapped in local minima. If it is too large, that may result in a wild divergent oscillating behaviour, and an increase in *E* liable to overshoot the minimum.

♦ **number of training cycles.**

Usually this depends on the learning rate and the number of hidden units. Training should stop when the error has converged to the global minimum. One possible solution to overcome the case of overfitting, is to add more training patterns as the training set may not be representative of the speech data space; or use less hidden units to avoid overtraining.

We combined these factors in the following training scheme to achieve good generalisation without excess of time required for learning the patterns.

1. Start training with a small number of hidden units, **[66]**,

   e.g. $\sqrt{InputUnits * OutputUnits}$

2. Start learning with a big learning rate and a small number of steps then gradually decrease learning rate and increase number of steps.
3. Record performance on the training and testing set.
4. Increase e.g. double the number of hidden units.
5. Repeat from (2) until error is acceptably small, or improvement is negligible.

4.6.3 ANN techniques employed

Classification and mapping has to be controlled from *OLTK*, and depends on the training and testing qualities of the particular *ANN* we use during run-time. This can lead to the problems which we saw in §**3.1.2-I-B**. For these reasons we extended an in-house piece of software for the training and testing of two types of *ANN*; in this way we could incorporate parts of it in *OLTK* by making certain routines to be compatible with the I/O processing of *OLTK*. The loading routine of our program is clever enough to build the already trained *ANN* simply by reading the filename of the weights. The first type of *ANN* is a posterior probabilities feedforward net (PPANN, §**Appendix-III-G**) that we use as a classifier; and the second is a sigmoid, mean sum-squared error feedforward net (SMANN, §**Appendix-III-F**) that we use to learn the mapping function from the cepstral domain to the 2D domain.

The software is written solely in C++ and makes use of a high performance floating point matrix/vector library, 'fltvec', developed by the Realisation Group at the International Computer Science Institute (ICSI). The algorithm used for training implements an 'on-line' mode which converges faster than the 'batch' mode and can more easily avoid local minima, **[52]**. In *on-line* mode for each *epoch*, the weights are updated after each training pattern is presented. Then we calculate the error for both the training and testing set. If the user-defined number of cycles is reached, or the error is lower than a minimum, training is stopped. Time elapsed since the beginning of training is estimated, and the weights are saved in a file. We can continue the training, if we wish, by reading the weights file. Typical durations for the data sets we have experimented with are from 1h to 4h on a Pentium 120MHz and with a minimum load of other processes.

I The a posteriori probabilities artificial neural network (PPANN)

It has been shown that the task of estimating the probability of the acoustic input vector belonging to a particular output class leads to a feed-forward *ANN* with softmax activations in conjunction with a cross-entropy training criterion, **[10]**. This combination increases the discriminability of the classifier, as it forces the target class to have an output probability close to one, and the other output classes to have probabilities close to zero, so that error is small. Kershaw, **[52]**, observed that *"this means that the training actively increases the likelihood of the correct class, while simultaneously decreasing the likelihood of the incorrect classes. This is different from conventional HMM training, where only the likelihood of the correct model (or state) is increased"*.

The algorithm for the implementation of the network follows the practical and theoretical consideration we have already discussed. In addition, we have merged a coding scheme in the weights file, where each 'one-out-of-n' binary encoded class can also be represented with a phonetic symbol. In this way the *ANN* is able to understand the format of the speech samples inside the data files (**§Appendix-III-E, Example 1**). The main role of *PPANN* in *OLTK* is for evaluation purposes. It has been used both in the real-time scoring display (**§5.3.2-C, §Appendix-II-C-b)** and in the frame play-back tool (**§5.3.2-B,** §**Appendix-II-C -b)**.

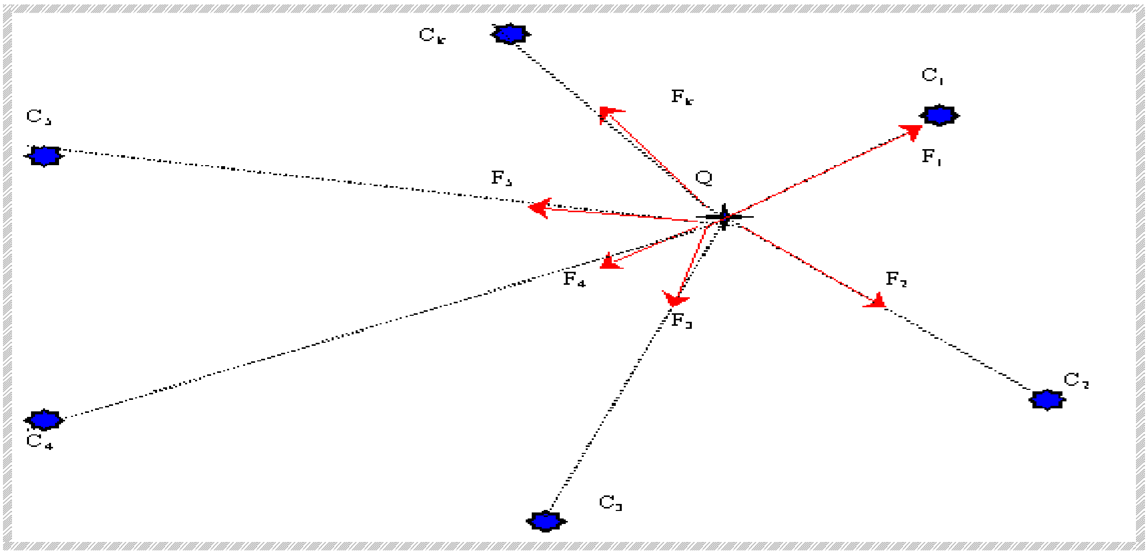


Figure 4-6 *Illustration of the central force model.*

The outputs of the *PPANN* have also been used to drive the mapping on the 2D display according to a central force model (**Figure 4-6**). We assume that there is a force, $\mathbf{F_i}$, exerted at any 2D point, **Q**, from any of the centroids, $\mathbf{C_i}$, of the sound clusters.

This force is proportional to the distance from the centroid, **d_i** , and the posterior, **P(C_i | Q)**. At any time, the system is in equilibrium, so the total force on **Q** is zero.

$$\Sigma F = 0 \Rightarrow \begin{array}{l} \Sigma F_X = 0 \Rightarrow \sum_{i=1}^{k} P(C_i \mid Q) * (X_Q - X_{C_i}) = 0 \\ \Sigma F_Y = 0 \Rightarrow \sum_{i=1}^{k} P(C_i \mid Q) * (Y_Q - Y_{C_i}) = 0 \end{array}$$

Where (X_Q,Y_Q) are the co-ordinates of the **Q** point and (X_i ,Y_i) are the co-ordinates of the **C_i** centroid.

The effect of the central model on the mapping of any cepstral vector that belongs to a specific sound class of the map is to plot most of the points at the centroid of this class. This is because usually the outputs of the classifier are all zero except for the target output one, which is usually one.

II The non-linear 2D fixed centroids mapping technique

The second feedforward *ANN* used in *OLTK* is the *SMANN.* This has been exclusively designed as a mapping function from the cepstral domain to the 2D domain of our display. The main characteristics of this mapping are the non-linearity and the fixing of the targets. In particular we create the patterns for training *SMANN* in such a way that all the samples of a sound class are mapped on a single 2D point. This point is going to be our 2D fixed centroid and the mean of a bivariate normal distribution. The clusters of points and the associated variances of the classes are constructed once we test *SMANN* with a codebook vector set, or with some other appropriate testing set (**Figure Appendices-7**).

A point to emphasise here is the calculation of the error. The mean sum-squared error (MSSE) is easily computed by summing the squared cepstral coefficient differences between what a predicted output should be and what it actually is, and then dividing by the number of cepstral coefficients. The squaring emphasises large errors. According to Masters, *"if the network is attempting to determine the presence of a particular signal pattern in a time series, the mean square error says nothing about the likelihood of missing the pattern if it is present, or falsely detecting it when it is not present"*, **[66]**. This makes it even more necessary to include the scope of the sensitivity prior to the mapping of the input vector. More reasons are explained in the following subsection.

4.6.4 Review of the out of training space problem (OOTS)

It should now be clear to the reader that the *OOTS* problem is due to our trust that the trained neural network will be able to interpolate between training patterns when it encounters unseen input cases. Of course it does matter how similar the input patterns are to those represented in the training set. In general, the more dissimilar these patterns are, the more doubtful we are about the integrity of the output from the net, and consequently about the result in mapping and evaluation of the utterance.

I Theoretical point of view

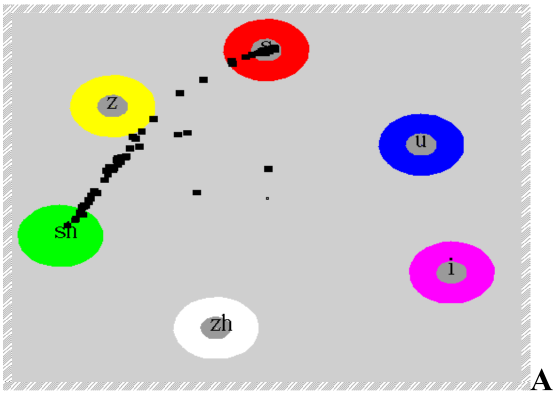
A term that is closely related with *OOTS* is the 'curse of dimensionality', **[10]**. If we consider the *ANN* as a mapping function from an input space to an output space, *ANN* needs to cover every part of its input space in order to know how that part of the space should be mapped. Since we are forced to work only with a subset of sounds, only this specific portion of the whole speech space can be mapped consistently. Two important observations are made here by Bishop, **[10]**: first, there could be a lower dimensional space where our data points can be restricted, assuming that features are generally correlated in some way; and, second, we hope that *"the value of the output variables will not change arbitrarily from one region of input space to another, but will typically vary smoothly as a function of the input variables. Thus, it is possible to infer the values of the output variables at intermediate points, where no data is available, by a process similar to interpolation"*. In particular, speech patterns that are outside the range of the training set, also called outliers, require extrapolation.

This leads to another important condition for good generalisation, which is smoothness. Masters observes, **[66]**, *"this implies a smooth transition between training cases"*. In other words, a small change in the input vector should produce a small change in the output vector. To conclude, smoothness and interpolation are the keys for a good mapping and an appropriate evaluation. What we have not examined so far is the relationship between the acoustics of the sound and the varying position of the articulators.

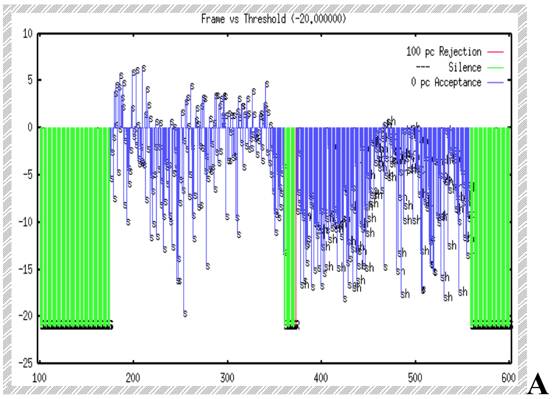
II Practical point of view

Requirements {C}, {D}, and {E} of the OLT method emphasise the fact that the sound models, represented on the map as cluster targets, should be linked with specific articulatory configurations, and that consistent changes in place and manner of articulation should produce consistent changes in the display. For that purpose, let us take a specific map to examine: the *iso2vo4fr* (**Figure Appendices-50**) will be our paradigm. From the models represented we chose to play with two sounds, the /i/ and the /s/, (for further similar examples and results see **[39]**).

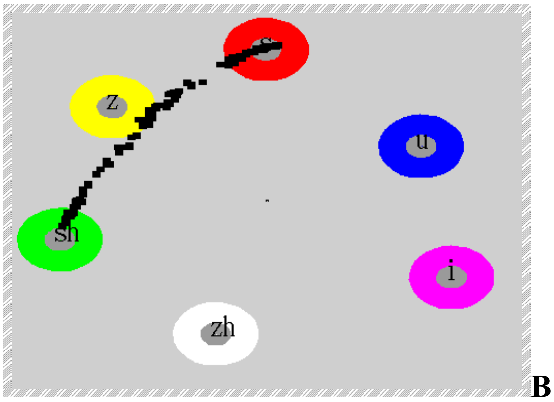
First, we have to check what particular articulatory configurations maximise the acceptance of our input. This can be easily done by setting the sensitivity to high values, (-2), and testing several productions of the /i/ and /s/ sound. In this case we found that a high, front position for /i/, and a blade raised very close to the alveolar ridge place of /s/ groove, gave high scores. Therefore these configurations should be considered the norm of the models. The next thing to try is to vary smoothly the articulators, gradually moving away from the norm of the sound, and check the effect on the mapping and evaluation of the utterance. To correlate better the change in place and manner of articulation and positions on the map we attempted to vary only one articulatory feature, as long as that was possible.


s
z
u
sh
i
zh

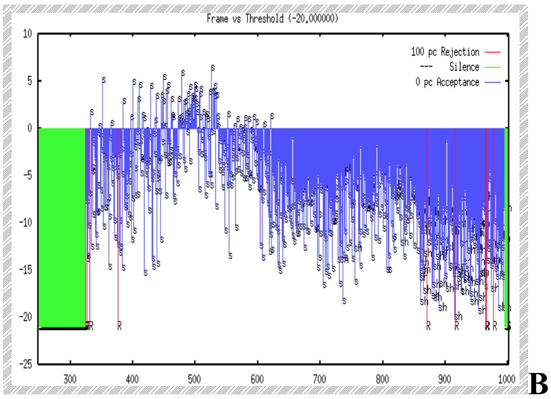
A


Frame vs Threshold (-20.000000)
100 pc Rejection
--- Silence
0 pc Acceptance
10
5
0
-5
-10
-15
-20
-25
100
200
300
400
500
600

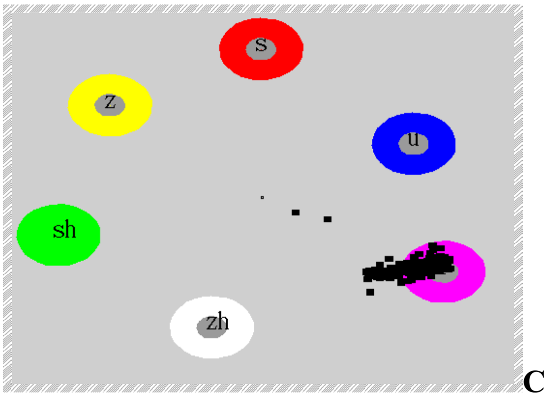
A


s
z
u
sh
i
zh

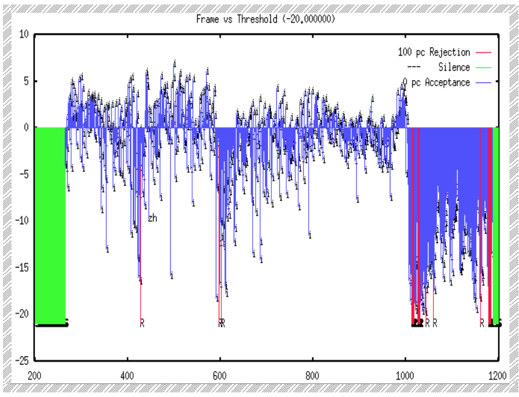
B


Frame vs Threshold (-20.000000)
100 pc Rejection
--- Silence
0 pc Acceptance
10
5
0
-5
-10
-15
-20
-25
300
400
500
600
700
800
900
1000

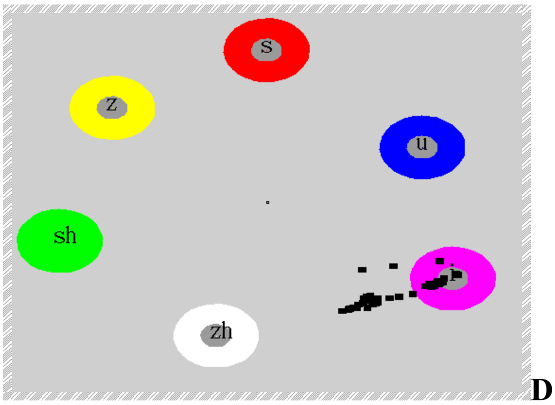
B


s
z
u
sh
i
zh

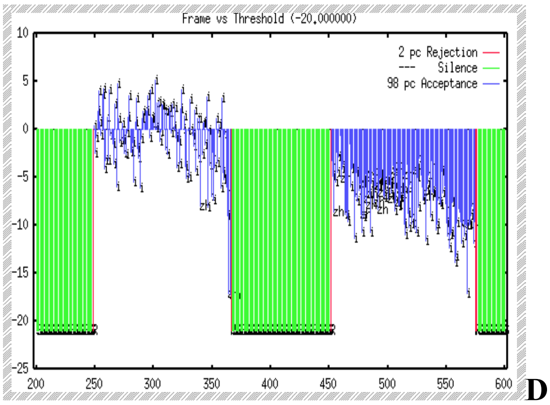
C


Frame vs Threshold (-20.000000)
100 pc Rejection
--- Silence
0 pc Acceptance
10
5
0
-5
-10
-15
-20
-25
200
400
600
800
1000
1200

C


s
z
u
sh
i
zh

D


Frame vs Threshold (-20.000000)
2 pc Rejection
--- Silence
98 pc Acceptance
10
5
0
-5
-10
-15
-20
-25
200
250
300
350
400
450
500
550
600

D

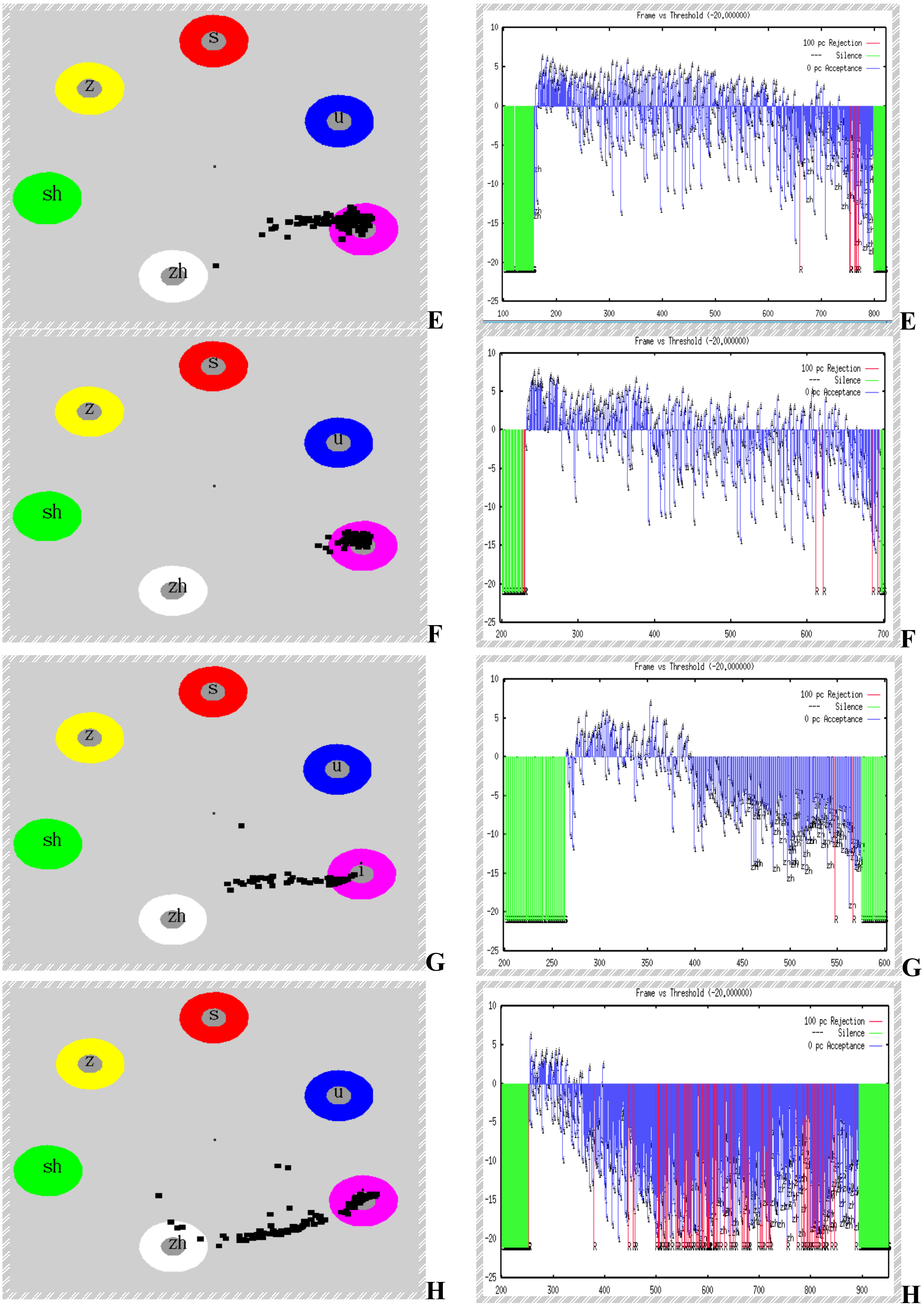


Figure 4-7 *Mapping and evaluation while varying articulatory positions. (A) /s/ - lip rounding, (B) /s/- tongue retraction, (C) /i/-raising pitch, (D) /i/-nasality, (E) /i/-lip rounding, (F) /i/-high to low (broadening oral cavity), (G) /i/-front to back (tongue retraction), (H) /i/-(narrowing oral cavity).*

A. Lip rounding and tongue retraction on /s/ sound

In **Figure 4-7-A**, we see the effect of lip rounding. If we ignore for the moment the /z/ cluster, clearly the unseen, from the *PPANN*, extreme rounded /s/ has been classified and mapped in the area of the /ʃ/ sound. Perceptually speaking, the effect of lip rounding can never result in a /ʃ/ sound; so the mapping can be characterised as inconsistent in that case. On the contrary the effect of tongue retraction on the same sound produces a consistent mapping. One can verify indeed that in retracting the tongue from a /s/ configuration (**Figure 4-7-B)** we can end up in a /ʃ/, as the groove position has been moved close to the palette.

B. Raising the pitch of /i/ sound

It is pitch that we claim to separate from the supraglottal vocal tract with cepstral analysis. Therefore we uttered the /i/ sound varying only the pitch to check the influence. Examining the acceptance of the utterance (**Figure 4-7-C)** we notice that it has remained at the same level except the last part after 1000msec where pitch increased most. Nevertheless this has not affected significantly the mapping in which it looks as though all points are close to the centroid of the /i/ cluster: this is another case where map behaved consistently.

C. Nasal coupling on the /i/ sound

A source that cepstral analysis does not take into account is nasal coupling. The effect can be seen in **Figure 4-7-D**. Although the frames with nasality have been accepted with relatively high values, the mapping of nasality is out of the /i/ cluster area; therefore the mapping can still be characterised as consistent.

D. The /i/ sound with lip rounding

In **Figure 4-7-E**, we can notice the effect of lip rounding. Most of the rounding is mapped inside the cluster of /i/ and the extreme cases are plotted in positions close to the cluster but are subject to rejection for high levels of sensitivity. Despite this fact the mapping is near or inside the area of /i/, and many frames score high, as was expected; the problem is that if we do need to differentiate and create contrast for rounded configurations of /i/, this is impossible with the current modelling.

E. High to low variation of /i/ sound

A more obvious similar problem results in going from a high /i/ to a low /i/ by broadening the oral cavity (**Figure 4-7-F)**. As we can see on the graph of the *HDF* metric, distances gradually start increasing, but all the plotting remains concentrated on the centre of the cluster. Therefore as in the previous case, we can characterise mapping as reasonable, but not consistent, since increasing the sensitivity will result in rejection of samples that perceptually can be judged very close to the model.

F. Front to back variation and narrowing the oral cavity for the /i/ sound

Two cases similar to those in **Figure 4-7-A,B** can be seen in **Figure 4-7-G,H**. The first (**Figure 4-7-G**) shows the effect of going from a front /i/ to a back /i/ by retracting our tongue. In contradistinction to the tongue retraction for the /s/ sound, here mapping is inconsistent, as moving the blade of the tongue towards back /i/ sound can be mapped in the /ʒ/ area. The second (**Figure 4-7-H)** shows the visual effects by narrowing the oral cavity. This is achieved by gradually creating a stricture of the tongue with the palate. The effect is to move along to the /ʒ/ cluster as the quality of the /i/ sound perceptually resembles that of the /ʒ/. Therefore mapping is consistent.

III Summary for the OOTS problem

Variation	**Consistency**	**Reason**
/s/ lip rounding	<A>-NO	It may be plotted on the /ʃ/ area.
/s/ tongue retraction	<B>-YES	Good correlation with 2D positions.
/i/ raising pitch	<C>-YES	Does not affect mapping unless too high.
/i/ nasality	<D>-YES	Plotted outside the cluster area but not far away.
/i/ lip rounding	<E>-NO	It may be rejected although close to the norm
/i/ broadening cavity	<F>-NO	It may be rejected although close to the norm
/i/ tongue retraction	<G>-NO	It may be plotted on the /ʒ/ area
/i/ narrowing cavity	<H>-YES	Good correlation with 2D positions.

Table 4-3 *Summary of results concerning the consistency of mapping.*

We tried to vary the articulators in a smooth manner so that the resulting acoustic features differ from the norm model according to a varying distance. There are cases where the network not only learns to map sounds which fit well in the modelled distribution, sufficiently close to the corresponding 2D area, but also learns to interpolate from one distribution to another, attempting to map these unseen samples in the space between one 2D cluster and another, <B> and <H>. Despite that fact, mapping inconsistency may occur in the 2D space between the target models by varying articulation in a manner which does not lead from one sound to another, <A> and <G>. This is mainly due to the fact that we ask the net to learn and map at a single point all the variant allophonic forms of a phoneme without giving any extra information about these deviancies. As an example consider all the <A>, <E>, <F>, <G> cases and especially lip rounding. The attempted extrapolation here leads to wrong results and inconsistent mapping.

A last comment is that forcing the net to map all phoneme variants to a single point does not help the smoothing of the mapping function. Practically this means that even if the input is accepted it will make no difference whether the change of a certain articulatory feature is small or large, as it will be plotted around the same point which is the centroid of the sound cluster, <E> and <F>.

4.7 Visualisation

It is time to return in the *SpeakaLook* cycle (**Figure 4-1**) and see how *OLTK* turns the speech events into visual events on the display.

4.7.1 Visual processing

There are four inputs to the visual processing engine of *OLTK*. Two of them are simply flags for silence detection and rejection. The silence state is reflected as an initial position during animation and it is also depicted on the face of a clown. Rejection simply denotes the absence of any animation apart from the change of the clown's facial expression, but affects of course the scoring of utterance. The 2D pair of co-ordinates marks the point where the mobile or a hit should appear, while the posterior probabilities are used in the scoring system to calculate how the percentage of the utterance accepted is classified.

4.7.2 *Visual feedback*

Therefore we can distinguish between, on the one hand, the animation that forms the qualitative, navigational feedback requirement and, on the other hand, the scores that satisfy the quantitative, evaluative approach in speech training. It is important to emphasise that both scores and animation occur in real-time.

4.7.3 *Reinforcement*

Each type of feedback, animation or scores, can contribute to the reinforcement requirement of speech training. The clients can be self-motivated either by trying to beat their high score, or by attempting to drive the mobile through certain areas to reach target positions, or by means of both. A reward in that case can have a huge impact on the performance, especially for the younger students. *OLTK* provides a simple reward once a certain score threshold is exceeded.

Reinforcement is the last step on the *SpeakaLook* cycle. Once the goal to reach, and the means to achieve it, have been fully appreciated and realised by the trainee, repetition should lead naturally to the acquisition of the missing speech skills.

4.8 A review of the implementation of OLTK

We end this chapter with a brief discussion about the most important and influencial issues concerning the software and hardware implementation of OLTK.

4.8.1 *Hardware implementation issues*

Speech processing can be a particularly demanding task for a computer. The real-time issues, the high quality of speech and the graphics animation require a powerful system. On the other hand we considered the cost of the equipment and tried to build the system with non-specialised hardware parts. Taking into account also mobility issues, we built the software on a laptop computer so the whole system is portable. The processor used was a Pentium 120MHz for the laptop and a Pentium 133MHz for the desktop system. Both computers were running a Linux operating system and were equipped with 48Mb of RAM and 2Mb video RAM for the SVGA cards. The hard-disk capacity was about 1Gb. The sound cards in both computers were Sound Blaster compatible. When OLTK runs, and recording starts, almost all the memory and processing resources of the system are allocated to the application. Of course today processing power has been tripled and perhaps the brute force of the processor is sufficient to run the OLTK without any clever programming tricks or techniques to increase the efficiency of the software, but by the time the program has been developed and with the existing equipment the algorithm has been modified many times in order that the update of the visual events on the display may be immediate.

4.8.2 Software implementation issues

The software programming design of *OLTK* has been particularly intensive and demanding. The program currently employs around 6000 lines of pure C++ code. Object orientation has been one of the most influential trends in programming over the past decade or so. Speed, coherence, better organisation of programming structures, and protection from side-effect errors, are some of the powerful features of the language. In addition, we used two extra libraries, one for graphics (EZWGL), **[109]**, and one for data structures (LEDA), **[71]**, to strengthen and expand the power of the language in the respective areas.

In order to appreciate the code design, we will refer briefly to the following most important programming aspects.

I Different programming states for callback events

Depending on the user action and the state of the program execution, there are different responses from the system that take place once the request has been processed. In our application we call these specific responses, 'main call-back events'. There are three different states internally implemented; two of them are for editing the *elements* of the map, *phones* and clusters (**§Appendix-II-B-c)**. The other is devoted to the recording/playback mode (**§Appendix-II-C**).

II Interprocess communication

One can certainly argue that one of the advantages of Unix over other operating systems (like Windows for example) is the superiority in running multiple processes. Indeed, the core of Unix has been designed with a multitasking, multiprocessing philosophy. If we consider also the relatively easy programming of the sound hardware and the sophisticated interprocess communication mechanisms of ‘pipe’, we have the main reasons why we selected to build our system on Linux a Unix clone.

The *pipe* is the simplest synchronised way of passing data from one process to another. We can envision a *pipe* as a conveyor belt. Data is written at one end of the *pipe* and is read from the other in a first-in-first-out (fifo) manner. We can implement the *pipe* in two ways: ‘unnamed’ *pipe* and ‘named’ *pipe* or (FIFO). A named pipe has the additional benefit that an unrelated process can use the *pipe* file which is a directory entry. *OLTK* software can make use of the *pipe* mechanism to implement the *SpeakaLook* cycle (**Figure 4-1**) with either a sequence of ‘unnamed’ *pipes*, ‘pipeline’, or both a *pipeline* and a *FIFO*. The scope behind the *pipe* implementation was to build certain parts of the cycle as independent processes. These processes can be tested independently of *OLTK* to trace for errors or check their result and performance. Each process can be connected with the next one in the *pipeline* with a *pipe*; so the output of one process, writes to standard output (stdout), can become the input of the other, reads from the standard input (stdin). *OLTK* generates the *pipeline*, reads the output of the final process and destroys the *pipeline* when not needed. Alternatively, *OLTK* creates a ‘child’ process to run in parallel *(forking)* and a *FIFO* to read the output of the *forked* process. In the meantime the *forked* process creates the *pipeline* and writes its output to the *FIFO* file that has opened for writing. As we are going to see, the real-time processing of *OLTK* (§**4.8.2-III**) is based exactly on this description.

The interprocess design makes OLTK architecture flexible, as any program that satisfies the I/O requirements can become part of the cycle. In addition it makes OLTK source code better structured and better organised.

III The real-time processing cycle

This has been the centre in our design philosophy in order to fulfil an important requirement {A}. The programs the real-time processing depends on are the parts of the *pipeline*. These are the *rec/playsample*, *sildetect*, and *HCodeRT* (§**Appendix-III-I)**. In §**II** we discussed two ways of implementing the *SpeakaLook* cycle. In both algorithms the *pipeline* is created either from the *forked child* process or directly from *OLTK*. Similarly, the classification and mapping can be performed either from the *child* process or directly from *OLTK* and the output of the *pipeline* is read either from the *FIFO* file or directly from *OLTK*. An advantage of the first method over the second is that, if the *forking* process is executed on a different machine, it will speed up the processing abilities of the whole system.

In order to measure the performance of the system we had to calculate the elapsed time for each processing cycle during recording or playback. A shortened version of the algorithm for the loop is as follows.

```
While (not end of recording/playback)
(1) Process the frame object
If not in silence period
        Calculate HDF distance
        If frame is not rejected
                Calculate posteriors
                Calculate 2D co-ordinates
Append or update the frame
(2) Animate the frame
Average every k frames
        (2a) Draw animation
        (2b) Update scores
(3) Check  the events queue
```

The *pipeline* and the statistical calculations have been the subject of discussion in the previous sections. An important point to mention is that the silence frames of speech are only detected not processed; so that reduces the overload of the CPU. We now turn our attention in another operation which requires much processing, the drawing of animation. Although our graphics library supports double buffering for drawing, we chose to draw directly on the front buffer, as it is only the portion of background under the old position of the *sprite* that needs to be restored once the *sprite* is moved to a new position. Thanks to the support of the graphics library on pixel operations, we performed a redraw-erase pixels cycle for the size of the mobile only. Even with such minimal animation requirements, the computing power by the time we were experimenting was just enough to perform the 10msec minimum redraw-erase cycle. We even had to use some special image processing to cut and shape the mobile to the absolute minimum size we could get. For the 'snake' animation we applied the XOR animation method. This is incomparably faster for consecutively drawing and erasing the object, as it does not require a previously read operation to store the original background contents. The price to pay is that the resulting image of the *sprite* depends on the background.

Another improvement in the speed of processing came from the algorithm on drawing the animation; actually there have been lots of modifications in general to make the code more compact and to avoid repetitions. An important modification came from the fact that certain visual events, like drawing the clown, or plotting the *sprite* on the initial position once silence is detected, take place only at the beginning of each state. Therefore the relevant parts of the code are executed once only after the initiating of each state.

To check whether the system performs in real time or not, we created a timer to measure the actual time elapsed. The timer is turned on before starting the real time loop, and off once recording/playback is finished. The number of animation cycles performed, $\mathbf{c_e}$, and the expected time duration, $\mathbf{t_e}$ **,** is checked, and contrasted against the actual time elapsed, $\mathbf{t_a}$, during that period. If the system runs on real-time the time measurements should coincide, $\mathbf{t_e = t_a}$ **,** otherwise in the general case $\mathbf{t_e < t_a}$ the system experiences delays. Taking into account that the frame processing is every 10msec, the expected real-time frequency is 100Hz and the actual animation frequency $\mathbf{v_a}$ is given by the formula :

$$v_a = \frac{c_a}{t_a} = \frac{c_e}{t_a} = \frac{t_e \cdot v_e}{t_a} = 100 \cdot \frac{t_e}{t_a} Hz \text{ and } t_e \leq t_a \Rightarrow v_a \leq 100 Hz$$

Typical measurements in some of the previously considered real-time optimisations can be seen in **Table 4-4 :**

Case examined	t_e	t_a	**Delay**	v_a
[A] Heavy loaded CPU	60sec	90sec	30sec	67.6Hz
[B] Size of sprite 100x86	60sec	85sec	25sec	70.6Hz
[C] Size of sprite 50x43	60sec	62sec	2sec	96.7Hz
[D] Drawing points	60sec	60sec	0sec	100.0Hz
[F] Unnamed pipe	60sec	60sec	0sec	100.0Hz
[G] FIFO	60sec	60sec	0sec	100.0Hz

Table 4-4 *Estimating real-time animation frequency for different cases*

In cases [B]-[G] the CPU load is considered minimal and the resources of the computer have been mostly allocated to *OLTK*. Also running the program with either unnamed pipe or *FIFO* we assume that the forked process is created on the same machine.

IV Recording/Playback program

We had to develop our own home made recording program (rec) as most of the existing recording programs are unnecessary complex and difficult to modify for our purpose. The *rec* program opens the digital signal processing (dsp) device for reading, and can write samples both to *stdout* for the pipeline, and to a specified temporary file. The last is a recording buffer we need to play-back the data. The *rec* program uses the I/O control device (ioctl) to set the bits, rate, channels and buffer size of the *dsp*. Our *pipeline* requires that we 'fflush' the buffers, which means the buffered data are forced to be written to the given output. Likewise, our play-back program, 'playsample', reads samples from a specified file, e.g. the recording buffer, and opens both *dsp* and *stdout* device for writing.

V Software implementation of cepstral analysis

In OLTK software implementation we use a modified version of *HCode*, (*HCodeRT*, **§Appendix-III-B**), that accepts input data from *stdin* and can write directly to *stdout* according to the requirements of the *pipeline*. Moreover we calculate the normalised log energy of the frame in real-time and provide extra arguments to set the minimum and maximum log energy so that we can normalise and scale the energy. The *MINSHORT* sample from *sildetect* process is also accepted from *HCodeRT* as a special signal to indicate the start of silence period. It is regarded as a frame where *MFCC* and energy are all zero. The zero vector of *HCodeRT* is the new indication mark for silence that is passed to the next process. An analytic example of the parameters we used and a run-time output of the processing can be found in **§Appendix-III-B**.

VI Flexibility

OLTK is a very flexible piece of software. This is mainly due to the independence from various other processes that are running in parallel with the control mechanisms of the interface and the presentation of graphics. In §**II** we talked about the *SpeakaLook* real-time cycle and highlighted how important it is for other programs to communicate with *OLTK* with appropriate I/O mechanisms. We also explicitly referred to the importance for independence from the type and structure of ANN to use (§**4.6.3)**.

VII Generality

We designed the code for *OLTK* to be as general as possible. That means that the layout of various displays is dynamic, depending on the number of sound classes in our data set. The map is created with a variable number of clusters, and different colours and labels are assigned. The evaluation display and frame playback tool are also designed to include a variable number of classes.

VIII Scalability

The main window of *OLTK* where the map appears is scalable. Practically this is a very convenient feature as the user can resize the window to fit in with other displays at one screen. It is even possible to run simultaneously two copies of the *OLTK* program from the same or different machines, in one screen for comparison purposes.

IX Fault tolerance

Special care has been given to detect and avoid the run-time errors that may occur because of accidental misuse of the interface functions. These include, errors during the loading of a map, and errors in pressing buttons or activating options that may crash the program.

X Maintainability, code readability

The size of code is already big enough to require efficient management. Readability and the associated maintainability are important factors once the program grows to a significant size. Thankfully, the code has already been redesigned from the old version of OLT completely from scratch to attain those qualities and laid to expanding. The current state is not satisfactory either: because of time limitations and user priorities we often sacrificed maintainability and readability.

Chapter 5

There were thus two things which the Saviour did for us by becoming Man. He banished death from us and made us anew; and, invisible and imperceptible as in Himself He is, He became visible through His works and revealed Himself as the Word of the Father, the Ruler and King of the whole creation (3:16).

St. Athanasius-On the incarnation (Translation-C.S.Lewis)

The speech training perspective of OLTK

In this chapter we discuss all the speech training aspects of the design and functioning of Optical Logo-Therapy Toolkit (OLTK). The efficiency of our application is much related to what we called effective visual feedback (*EVF*), and depends on the requirements we set out in §**2.2** (**Table 2-1**). Therefore the role of the programmer is to develop an understanding of all these factors and build appropriately the modules of the application, the tasks to complete, and the strategy to follow in order to enhance the role of the instructor and help the students to acquire the missing speech skills. As Katz notices, **[50]**,*"the quality of available software will improve as program developers become aware of the needs of the clinicians and as clinicians become discriminating consumers"*.

5.1 A user perspective on the interface

Although we have taken into account the needs of the instructors and the students, OLTK embodies tools for more technically oriented persons. At this stage of development this was both necessary and unavoidable. Necessary: since OLTK still serves as an experimental test-bed, some of the operations we describe, like editing the elements of the map, are there mainly for modifying and checking manually the transformation from acoustic vectors to points on the two dimensional plane. Unavoidable, because other operations, like the graphs for the evaluation of utterance, or frame-by-frame acceptance/rejection, are conceptually rather complicated for the non-expert. The last is due to the fact that we attempt to analyse in depth the articulation, and see where and how exactly the error in production occurred and what was the effect on neighbouring phonetic segments.

From the therapist's point of view, the key issues in the design of OLTK is simplicity and effective functioning. These two issues can conflict, because including all the control features and keys which the therapist needs can result in a rather complicated interface. The problem is further complicated if we take into account the different demands of other professionals who might use the application, like language instructors, or perhaps phoneticians.

For these reasons, although OLTK has a number of features that require technical expertise to handle them, we provided a minimum number of easily accessible options for the therapist to use in experiments with the patients.

5.2 Overview of OLTK functionality

OLTK is fully operational and accessible through a graphical user interface. All options of OLTK application are accessible through keystrokes or mouse-button clicks. The menu options have been designed in the same style that all graphical user interfaces follow nowadays. In that fashion certain type of buttons (§**Appendix-II-A-a)** serve a two-way purpose; first the user is able to activate or deactivate a certain feature of the application, and second it is possible to display or hide other essential components and control windows of OLTK. The last is necessary in order to avoid cluttering the working space and confusing the user. In this way the user can also choose to bring up only those windows necessary, depending on the training stage and the mode of operation.

In general we tried to provide options for both the instructor and the student. According to Katz, **[50]**, *"whenever possible, patients should be offered the opportunity to select options…for patients who may be confused by options, preset default values can be displayed …"*. Indeed, to simplify things we provided default options for both the instructor and the student (**Figure Appendices-2**); therefore the training can start without fiddling around with many settings. Moreover, access to and interaction with the various options and operations of the application (particularly for non-expert, non-technical persons) has to be effortless, attractive, and efficient. The design of the menu buttons and the splitting into groups serve those purposes.

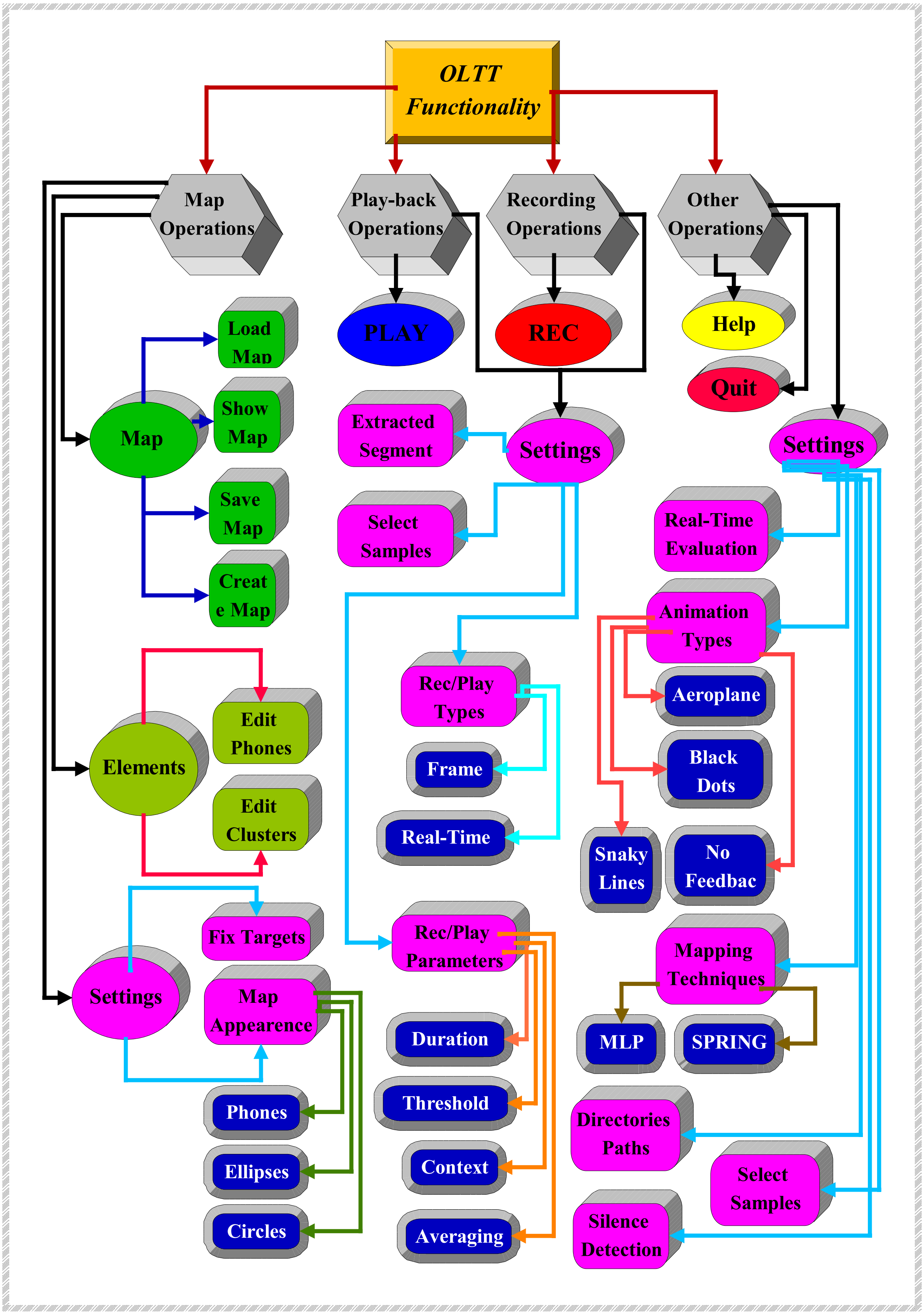


Figure 5-1 *OLTK functionality: Different shapes indicate the level of the options and different colour the different option groupings. Functions are grouped per column.*

We present a re-organised grouping of all the options available in OLTK in **Figure 5-1** according to four different operations of the program: 'Map Operations', 'Play-back Operations', 'Recording Operations', and 'Other Operations'. Under each one of these four headings we can see what menu options are related. The six ellipsoid shapes, each one with a different colour, indicate the main buttons available on the menu bar; the cubic shapes indicate what options are available under each pull-down main button; and the plaque style shapes indicate the options available in a deeper level of pull-down buttons. We can also notice the common options for both the play-back and recording operations of the program.

5.3 Practical justification of OLTK options

In the rest of this chapter we provide explanation only for those options of OLTK mostly used, and mosty important in the speech training; what is their role and how they can be handled effectively by the instructor. The rest of OLTK options are described in §**Appendix-II**. We can split these options into two groups; 'Map Options' and 'Recording/Playback Options' depended on the diagram of OLTK functionality we drew in **Figure 5-1**.

5.3.1 Map options

The design of and experimentation with phonetic maps in OLTK is of primary importance for the developer of the application, but also for the instructor. Phonetic maps form the heart of the application; therefore many options have been provided for map functionality. The map operations include three groups of options, 'Map Options', 'Elements', and 'Settings' (**Figure 5-1)**.

A. Loading process

Loading an existing map **(§Appendix-II-B-a)** is the first necessary action. The loading process of OLTK is relatively complicated, because there are many other files and routines related to a phonetic map. Therefore it was important to hide all these technical details from the non-technical person and simplify the function of loading. Once loading is finished the user can have access in all the other options of OLTK such as changing the appearance of the map.

B. Map appearance – Animation types

From one aspect, selecting different appearances for the map (§**Appendix-II-B-b)** gives both the instructor and the client more freedom on how to visualise the sound targets. The different graphics representations of the map appeal visually to the client. It is possible for the instructor to think about different task descriptions that fit with a particular representation by selecting one of the various animation types available in OLTK **(**§**Appendix-II-D-a)**. For example: the drawing of the map with boxes scattered and clustered in different areas can be used to instruct the client to black out the boxes with the *Black Dots* animation type (**Figure Appendices-7**), or fly the *Aeroplane* to a cloud drawn as an ellipse or a circle (**Figure Appendices-9**).

On the other hand, the transformation from multi-dimensions to two dimensions can be visually inspected (**Appendix-II-B-c)**. This makes it possible to spot the 2D outliers, examine them and recreate the map with less variance and better clustering (**Figure Appendices-9, Figure Appendices-10**).

5.3.2 Recording/Playback options

Speech training with OLTK is mostly based on the 'Play-back' and 'Recording' operations (**Figure 5-1**). In our interface design philosophy we tried to treat these operations concurrently; hence many options are shared, and the various visual events can be repeated with both modes of operation. The 'Real-Time', 'Rec/Play Parameters', 'Animation Types' and 'Mapping Techniques' options of *Settings* main button menu work are for both *PLAY* and *REC* functions. In practice this means that whatever action and result of a student captured during a recording session, *REC*, it can be reproduced identically with the accompanying audio and visual feedback, *PLAY*. It worth saying here that during the experiments with the therapist, children found it particularly amusing to start the recording or play-back of their speech on their own simply by pressing these coloured buttons with the mouse.

A. Real-Time audio-visual feedback

Real-Time means that audio-visual feedback is immediate. During recording, the temporal relationship of what is being spoken and what appears on the display is clear. The user can change in real-time the visual patterns on the map in a systematic way aiming to reach a particular configuration. Moreover, other kinds of feedback, such as indications for acceptance/rejection of speech input and evaluation **(**§**5.3.2-C)**, also happen in real-time. Similarly, during play-back the recorded speech is reproduced by the computer and all the events we mentioned can be seen again. In play-back mode, *Real-Time* is a particularly desirable feature of OLTK as the instructor can focus on problematic segments, and examine in more detail the speech of the client by repeating the utterance spoken. In addition, the client can understand better what went wrong by repeating his/her last effort and relating the audio feedback to the visual patterns of the display. This can be an effective training mode, as the trainer can comment on the production of the trainee while s/he listens and observes the events on the display, something which perhaps is more difficult to assimilate at the time of speech production. This is both because the trainee may concentrate a lot on how to articulate correctly the utterance rather than paying attention to the instructions, and because the speech of the instructor may cause problems in the speech recognition of the system, especially if the microphone is picking up the voice of the instructor.

B. Frame by frame analysis of speech

We ought to say that the control of the 'Frame Play-back Tool' (§**Appendix-II-C -b)** is at present particularly complicated for the instructor; therefore it has not been used in speech training. Nevertheless, it helped us in researching the mapping, in classification of certain segments of speech, and in extracting the parts of the utterance that we were most interested in. It can be combined excellently with other OLTK displays like the one at **Figure Appendices-20** to focus on particularly interesting segments of an utterance.

C. Real-Time evaluation

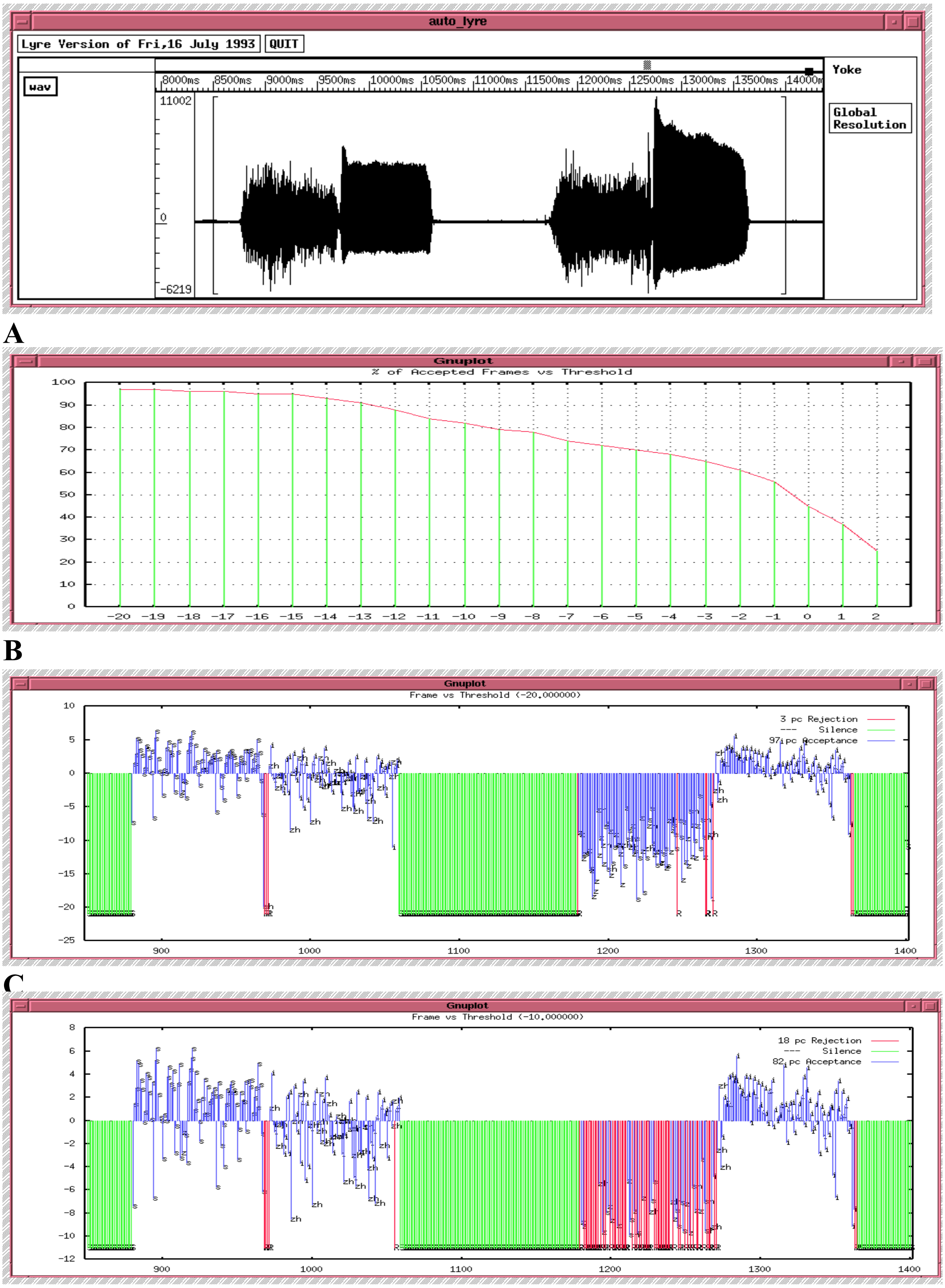
Requirement {F} for *EVF* states that a metric is required to measure speech quality. Our real-time evaluation display (§**Appendix-II-C-b)** provides evaluative feedback which is optional according to whether we set this display on or off with the check-button of *Settings* menu. Therefore it is possible for the instructor to make the student focus more on the navigational feedback of the map rather than paying too much attention on the scores gained.

In general the appearance and the design of the display is such that both the trainer and the trainee can easily grasp the score and relate it to the other graphics events of OLTK. It is also possible to use only this type of display for speech training on isolated sounds, instructing the client to raise sufficiently the height of the corresponding sound bar, and at the same time maintain a big score for acceptance. This aspect of OLTK reminds us of the speech training displays of *VSA*, (§**2.3.2-G, Figure 2-8**) and *VATA* (**§3.1.2-II-B, Figure 3-25**).

The scoring of the real-time evaluation has also been used to provide an extra form of reinforcement by rewarding the client for his/her efforts with stars drawn at the bottom of the screen (**Figure 5-3-B1,B3,C1,C3** and **Figure Appendices-27**). The star is drawn once a threshold level of sensitivity is reached.

D. Sensitivity

This is the option that makes the biggest impact in speech training with OLTK and the one most used by the therapist in our experiments. The 'sensitivity' is associated with the acceptance/rejection threshold (an in-depth theoretical and practical explanation is given in §**4.5.6-II** and §**4.6.4**). Most important sensitivity can be changed interactively during the recording or play-back operations. In this way the instructor can change the quality of visual feedback on-line as the student is experimenting with various articulatory gestures. The logic behind the slide-bar controlling the level of *sensitivity* **(Figure Appendices-18)** is quite simple. As its value is decreased, the system becomes more tolerant and the client's utterances can deviate substantially from the sound models present on the map. At this tolerance level, positive acceptance, classification and mapping are plentiful. However, more false-positive feedback may also occur (§**4.6.4**).



D

Figure 5-2 *Evaluation and frame-by-frame comparison between a normal and a lateralised /si/ for different sensitivity levels. (A) time waveform, (B) graph of acceptance vs sensitivity, (C) and (D) graphs of frame-by-frame acceptance/rejection vs distance and classification for sensitivity levels of (-20) and (-10) respectively. The red lines indicate rejection, the blue acceptance, and the green silence.*

At the beginning of speech training and particularly with *FAD* cases the instructor can set the sensitivity to low values to allow a broad interpretation of the client's response. When progress is made, and the instructor and student are confident on successful attempts, the *sensitivity* is gradually increased to encourage a more precise production. Thus a more accurate approximation of the phonetic model is required, which results hopefully in mapping on areas very close to the target models (see problems in §**6.2.4-IV**).

In order to see the effect of altering the level of *sensitivity* we have extracted a part of the speech signal recorded, using the F10 key and the context range of the *Frame Play-back Tool*. The utterances extracted are a normal /si/ and a lateralised /si/. In **Figure 5-2** we can see the waveform, (A), together with an F11 graph of the percentage of utterance accepted for each different sensitivity level, (B). Observing this graph one can find where to set the *sensitivity* level so that the client can easily attain a high score, e.g. 80% acceptance. In more detail now, for a sensitivity level of (-10) almost half of the frames of lateralised /s/ are not passing the threshold criterion, (D), while for (-20) they are mostly accepted, (C).

The above can be considered a microscopic analysis of the utterance. A macroscopic analysis is presented in **Figure 5-3**. First, examining the spectrum of each utterance with the *SFS* program (F8), **[47]**, we notice the extra formant of the lateral fricative around 3kHz, (A2), and its energy significantly lower than the normal fricative, (A1), in other frequency bands. Second, setting the level of sensitivity to (-10) and comparing the mapping between the normal, (B1), and the abnormal, (B3), we notice that the quality of the vowel /i/ is different, but -most important- that the lateralised /s/ has been plotted far away from the cluster of /s/ sound. A similar picture is produced with the sensitivity level at (-20): More points are plotted this time, but again the lateral fricative is completely out of target. A different visual presentation can also be seen with the *snake lines*, (D1) and (D2). Finally a quantitative comparison through the scoring displays, (B2), (B4) and (C2), (C4) shows that although the lateral fricative is almost totally accepted at (-20), it is mostly rejected (80%) at (-10), while the normal one scores above 95% acceptance in both cases.

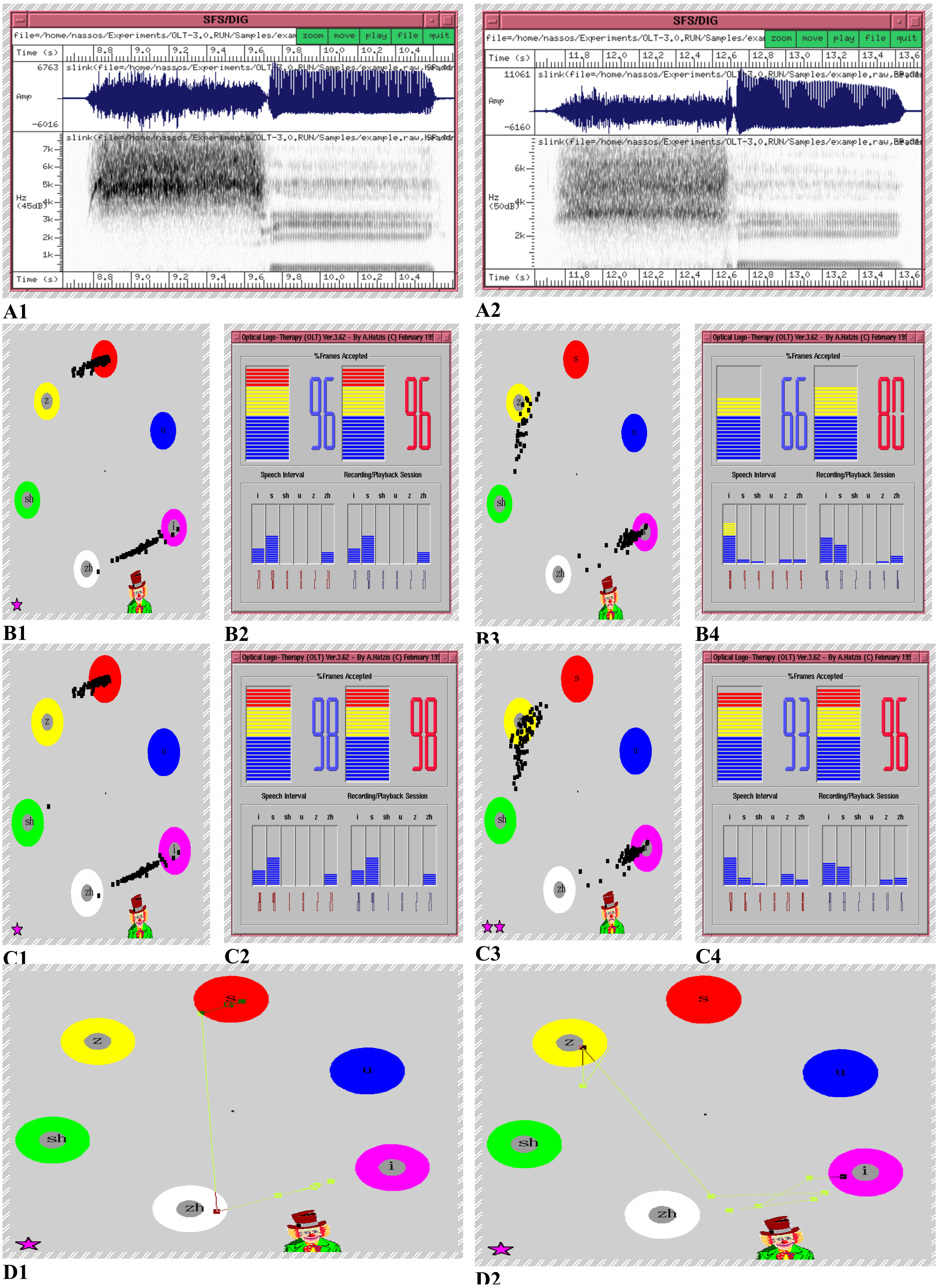


Figure 5-3 *Visual contrast and evaluation of a normal /si/ (left columns) against a lateralised /si/ (right columns) for different sensitivity levels, (-10, B) and (–20, C).*

Notice also that from the percentage of the utterance accepted, the fricative part is classified accurately only in the normal /si/ (50% /s/, B2). The percentage of the /ʒ/ is due to the vowel (27% /i/ and 21% /ʒ/, B2). However for the classification of the lateralised /si/ in (C4), the 93% acceptance breaks down to 50% /i/, and only 16% for /s/.

A last point to mention is that the *Sensitivity* option can be combined with the '-*Threshold'* parameter of the cluster (**Figure Appendices-13**) to tune the acceptance/rejection criterion of each sound category separately.

E. Duration

This is another commonly used feature of OLTK. The instructor can select either to start an unlimited recording session so that s/he can experiment with the speech production of the client, make tests, and offer directions and instructions on how to accomplish the tasks assigned, or check the button next to *Duration* label and set the slide-bar to a given number of seconds (**Figure Appendices-18)**. The last can be useful for setting time limits on a game like strategy, and reinforcing in this way the client's determination to reach his/her maximum performance on time. This can be beneficial for both the client and instructor in terms of effectiveness over the duration of speech-training sessions. A limited duration can also serve the purpose of saving recordings for future reference.

F. Averaging

Another less popular feature of OLTK is averaging. The *'Averaging'* slider (**Figure Appendices-18)** can set the number of frames to be averaged in terms of their x-y transformed co-ordinates. The visual effect produced is a kind of smoothed animation on the map when averaging the 2D positions of many frames; as we have noticed, this can produce a better aesthetic result for isolated sounds. On the other hand the flickering during animation increases for small values of averaging. But it suits much better the visual feedback for articulation of vowel-fricative segments, and the *Snaky Lines* are also better produced in this case. It also means that many more traces can be left on the map when the *Black Dots* option is selected. In general terms, although averaging is not critical to the behaviour of the system, the instructor can experiment interactively with this option to change significantly the quality of the visual feedback produced.

G. Context

The *'Context'* slide-bar varies the width of the segment when we examine an utterance with the *'Frame Play-back Tool'* (**Figure Appendices-16**). In other words it specifies the number of frames before and after the current frame we observe. In this way we can focus and examine the utterance in detail, and if we wish we can save the segment for further study or reference.

Chapter 6

> *Here, then, is the second reason why the* ***Logos*** *dwelt among us, namely that having proved His Godhead by His works, He might offer the sacrifice on behalf of all, surrendering His own temple to death in place of all, to settle man's account with death and free him from the primal transgression. In the same act also He showed Himself mightier than death, displaying His own body incorruptible as the first-fruits of the resurrection (4:20).*
>
> ***St. Athanasius-On the incarnation (Translation-C.S.Lewis)***

OLT in Action

In the previous chapter we have described analytically the real-time audio-visual feedback method of OLT and highlighted potential use in speech training. In this chapter we see how the different modules of OLTK have been applied in real conditions and we examine the application of OLT in two suitable, simple, but characteristic cases of speech training namely the FAD and AM.

6.1 The targets of the phonetic map in OLT

Phonetic maps with targets play a central role in OLT. For this reason, before we see how these are used in practice, we will attempt a more in depth theoretical analysis of these terms.

6.1.1 OLT phonemic representation

Considering the segmental nature of both the *AM* and *FAD* cases we examine, we notice that the problem is located in the English sibilant fricative sounds. The absence of distinction in the first and the lack of contrast in the second require a teaching approach that gradually will make the speaker abandon his/her abnormal sounds and acquire the skills to produce the normal sounds. Clearly this approach implies that we think in terms of the phonemes of the language. This is precisely because by definition a phoneme is an abstraction that serves to describe the contrastive or distinctive sounds within a language and conceptualise the phonological system. Actually the phoneme is a convenient link between the written and spoken forms of a language. We chose initially a phonemic representation for the map, because of the familiarity of clients with the alphabetic writing system of the English language. In the psycholinguistic model we described in **§1.2.2** a link exists between the orthographic representation and the phonological representation. That means that although in general *"the organisation of speech is in principle independent of orthography… it may well be true that knowledge of a writing system facilitates a certain analytical awareness of segments and structure"*, **[15]**.

6.1.2 From phonemes to targets

OLT maps short-time acoustics onto points in a 2D space associated with labels provided by the teacher, given training data annotated with these labels. Initially, these acoustic targets and the labels represent phonemes. Since here we deal only with fricatives and vowels, we can assume that at least potentially these sounds have a steady state; *"this stable state is assumed to include all the articulatory settings that best characterise the sound in question, and is referred to by phoneticians as a target… That is, the tongue, lips and jaw are meant to achieve- however briefly- a stable configuration, commonly called the target configuration"*, **[15]**. These sounds can be produced in isolation and prolonged, in that case *"it does make sense to speak of genuinely stable targets- at least potentially, for the stable target will not necessarily be observable in running speech"*, **[15]**. However, according to Ohala, *"due to non-linear mapping from articulation to aerodynamics and to acoustics there do exist near steady-states in these latter domains"*, **[72]**. Therefore, since we map the acoustics of certain stable articulations, we can expect that during the production of a syllable a normal speaker should pass near these target areas of the map. OLT provides audio-visual feedback to help identify and reach the targets. We consider that the ultimate function of these targets is to make the speaker focus on the relationship between the acoustics and stable configurations. It is the visual contrast that OLT creates between such stable configurations that is important for teaching purposes. It can claim to facilitate the connections between motor patterns (or articulatory gestures) and acoustics ultimately necessary for a production which is perceived as the correct phoneme.

6.1.3 Targets and coarticulation

However handy the concept of target may be speech cannot be considered as a series of static targets. The 'target-centred' view of phonetics is now generally accepted to be a simplification and *"less segment-based views of phonological representation and of speech production/perception have led to different ways of talking about coarticulation"*, **[44]**. Modern theories like action theory, **[28]**, and task dynamics, **[14]**, associate discrete gestures - vocal tract constrictions e.g. velic and glottal gestures, lingual gestures - that do not demand that the speaker attempts to reach static target positions. Instead they determine a dynamic modelling of a movement trajectory for each gesture variable, e.g. place and degree of constriction, controlled by the period of activation of the variables known as gestural score, **[40]**. The modelling for each gesture variable describes the behaviour of a mass connected to a spring and a damper. The resting or equilibrium position of the spring represents the target of an actively controlled gesture variable and is specified in the gestural score. Moreover coarticulation is treated more as a coproduction of various gestures. In this view, while OLT speech training to achieve acoustic targets do not claim to embody a theoretical explanation or model of speech production/perception it appears to have certain similarities with the task dynamic model we described. In particular, the mapping positions are dependent on the target models of OLT that act as attractors during the production of an utterance due to the neuro-mapping properties. It is also possible to vary a certain phonetic feature, gesture variable, e.g. place of lingual stricture, and observe a visual trajectory from one target location to another (**Figure 4-7)**. Even the central force model (**§4.6.3-I)** shares the philosophy of the spring/mass model of task dynamics. Despite these similarities the main idea is that modelling in OLT is such that it is simple, attractive, and comprehensive for the speaker, serving primarily the purpose of training by relating what is on the screen with his/her mental representation of speech and hence to his/her production/perception system (**§1.9**). It is also important to acknowledge the real-time visual feedback issues (**§4.8.2-III**) that require a fast pattern processing which is relatively more difficult to accomplish in dynamic modelling. As a last point to make, the view of coarticulation in OLT exists only for the transition between the targets that appear on the map. Since much of the training is limited to the contrast between isolated sounds and the targets of the map have been modelled from sustained phoneme production, coarticulation has a limited effect on the mapping of an utterance on the map. Nevertheless gliding from one sound to another as in consonant/vowel syllables can resemble some of the character of coarticulation.

6.2 Speech Therapy

Spatial distortions, like lisping, which frequently occur during abnormal fricative production are very common in children with functional articulation disorders [35]. There have been many similar studies based on EPG remediation with analytic and detailed descriptions of procedures and results **[16]**, **[29]**, **[46]**. Such studies offered us valuable information concerning the remediation schedule we followed with OLT.

6.2.1 The target problem

I Description of the speech disorder

Perhaps one of the most common and frequent of functional articulatory disorders is misarticulated sibilant fricatives. Hardcastle and Gibbon, **[35]**, uses the term 'spatial distortions' to refer to the abnormal articulatory configurations that occur with speech-disordered clients. The English alveolar and post-alveolar sibilant fricatives /s/, /z/, /ʃ/, /ʒ/ require a rapid, thin stream of air to pass across a sharp edge in order to produce a hiss, **[12]**. To narrow that egressive airstream we use the tongue so that air escapes along a central groove created with the palate. The closed teeth provide the sharp edge. As Howard states, **[46]**, *"Groove width and location are closely interrelated in distinguishing /s/, /z/ (narrow groove, anterior location) from /ʃ/, /ʒ/, (wider groove, more posterior location)"*. There is a slight difference here in the production of sibilant fricatives by many normal speakers depending on whether the tongue tip is behind the lower incisors and the blade of the tongue is raised to the alveolar ridge to make the groove or the tip of the tongue is grooved near the centre of the alveolar ridge, **[12]**. The common primary feature is the lingual-palate stricture and the opening for the air stream in the centre of the alveolar ridge. The common misarticulations that occur are the central lisp or frontal lisp (lingual protrusion or interdental), the recessive lisp, and the lateral lisp. In the first and the second case of lisping the air escapes from the central groove but the tip or blade of the tongue is placed either too far forward, frontal, or too far back, recessive, **[12]**. Lateral lisping means that the air escapes across one or both sides of the tongue rather than the centre **[12]**.

II Description of the clients

The identification of the type of disorder was the first step towards the completion of the arrangements for the experiment to take place. Then we discussed with the clinicians what population appearing with that type of disorder would be most appropriate for our experiments. We were advised that it would be easier to recruit children between ages of six and nine years old, as it appears that we can find many of them in the local area with that specific disorder. We recruited three school children, referred to here as ChildA, ChildB and ChildC, aged between five and eight from the case-loads of practising speech and language therapists in the Sheffield Speech and Language Therapy Service. All the children experienced difficulty with the articulation of sibilant fricatives but they had normal hearing, vision and cognitive ability and they did not have any other speech and language impairments and no other accompanying disabilities. The report of the speech therapist (§**Appendix-XI**) provides a brief outline of their individual patterns of difficulty with /s/, /z/ and /ʃ/.

III Etiological – Maintaining factors of the speech therapy case

It is a common practice in medicine, but it is even more important in speech therapy, that before any attempt for remediation of the speech disorder the clinician should examine thoroughly all the etiological or maintaining factors that cause the problem or have an influence on the development of it. *"These include sensory loss, particularly hearing and visual, structural and functional abnormalities of the vocal tract, cognitive ability, auditory discrimination or linguistic difficulties, medical factors, psycho-social problems, poor attention, motivation"*, [35]. It is in view of those factors that the speech therapist decided that the application of OLT has the potential for improvement of the client's speech (§**Appendix-**XI).

6.2.2 Experiment preparations

I Selection of normal subjects for modelling and comparison

In order to meet the needs for building our phonetic maps for our client group we decided to arrange a series of recordings of normal speaking children in local primary schools. We prepared a request form and sent it out to several schools asking for permission and arrangements to carry on with the recordings (§**Appendix-XII**). Soon we received positive replies and we picked up two schools to serve our purposes. We recorded in total eighteen children with ages from six to nine years old, half of them were male and half female. Before the recordings, we spoke with their teachers and made inquiries about the linguistic origins of their parents and about whether they noticed any abnormalities in their speech. Finally, we tried to select children with clear and good voices so to collect as good as possible samples.

II Normal Children Recordings

Speech Data Tool (§**Appendix-III-C**) was used here to control the acquisition of data. The corpus consisted of six sounds in isolation, three vowels /i/, /o/, /u/ and three consonants /s/, /ʃ/, /z/ as well as a list of words that contained those sounds (a sea, a zee, a sheep, a saw, a shore, a zorr, a zoo, a shoe, a suit). The duration of the isolated sounds was three to four seconds and the words were recorded five times. The recording format and equipment used was that described in §Error! Reference source not found.**,** **§**Error! Reference source not found.. The environmental conditions were good: recordings were made in a quite room with no external noise.

III Examination of the recording data

Three persons, among whom was a trained phonetician, listened to and judged the quality of recordings. The quality was found to be generally acceptable; but despite the many precautions we took during the recording time, several problems were encountered that were not noticeable at the time of recording. These include creaky voice, loudness, difficulty of sustaining the voice causing changes in pitch and sound quality. Also the sound quality was particularly unstable at the beginning and ending of the speech signal. Finally variation in each phoneme pronounced in isolation was also something we had to pay attention to. The problem was mostly located in sounds like /u/ and /o:/. We noticed that for /u/ there are two major variations one with lip rounding as the predominant feature and the other without lip rounding but with fronting approaching the /y/. Similarly the /o:/ sound for the female subjects was found to be close to /⊃/. The production of fricatives appeared to be more consistent and less variant for the population. There were only minor problems caused by casual dentalisation or retraction, and these were located and isolated during manual segmentation (§**4.5.2, §Appendix-III-D)**. This way we effectively discarded all the bad samples.

IV Building the phonetic maps

According to the analytical technical description we provided in sections §**4.5** and §**4.6** and the theoretical discussion in §**6.1** we modelled the targets of our maps from consistent, sustained, sound production. Using the set of commands in §**Appendix-III-H** we built two maps; one with three vowels and three sibilant fricatives (**Figure Appendices-41**, **Figure Appendices-42)** and another with three vowels and two sibilant fricatives (**Figure Appendices-43**, **Figure Appendices-44**). The datasets were collected from the recordings of the 18 normal children (§**6.2.2-I**). Three sets were created in total; two for the training and testing of the map and an LVQ one for modelling the targets. The accuracy of both the ANN employed for classification and mapping found to be satisfactory. The details of these configurations can be found in **Table Appendices-15** and **Table Appendices-16**. Finally the layout for each map was arranged in agreement with the therapist and was designed to suit the needs of training for the specific speech disorder of the client.

We tested the quality of our phonetic maps by trying recordings of two normal children. These children have not been recorded for the training of the map. We called for an official testing a male, five-years-old child. The therapist explained to the child how the software application works both by giving verbal instruction and by playing back previously recorded utterances. The level of comprehension was high and the child easily understood the task to accomplish. Pretty soon the child was enthusiastic with the use of OLTK and enjoyed the various drills that had been assigned to him. It was the first time since the development of the application that we received positive reactions by a child of that age. Then we performed the test on another child this time female of about the same age. Our judgement concerning the comprehension and enjoyment with the use of OLT was at the same level.

The testing utterances recorded included isolated sounds both similar and non-similar with the target models of the map, as well as some word-testing. With these we checked the level of acceptance, the classification, and the mapping on two dimensions. An analytical spreadsheet with the figures from that testing can be found in **Table Appendices-5, Table Appendices-6**. On average, the performance of the normal children on the maps was judged good enough to proceed with the training sessions of the abnormal children. The response of OLTK on the various recordings was reasonable and we could always identify a reason when the behaviour was not the one expected. In particular, referring to the aforementioned spreadsheet, we can explain certain surprising figures.

In the case of the isolated vowels, loudness plays a critical role. If the sound of the vowel was very loud OLTK level of acceptance was poor. Another reason for rejection of the vowels, especially the /o:/, was the variant way of speech production making it sound like an /ɑ/ or /⊃/. Similar problems appeared with frication. In contrast with the vowels case, if the loudness of frication was not above the silence detection limits, it could very easily be classed as silence. Here we also notice that due to the strong effort needed to make the fricative sounds and sustain them part of the sound could appear dentalised or palatalised which of course means poor acceptance and wrong classification. Finally testing OOTS sounds revealed the problem discussed thoroughly in §**4.6.4** and its importance became apparent during the speech therapy training sessions (**§6.2.4-IV)**. A characteristic example was the simulated dentalisation of tst-ChildA (**Table Appendices-5**, **Table Appendices-6**, row marked *'tst-ChildA-blow'*). The sounds were not rejected as it was expected but accepted with high scores and the 2D positions were mapped exactly on the target positions. This problem has been tackled with the use of individualised maps (**§6.2.4IV)**.

Finally, map '*ch-s-sh-i-u-o-lvq'* (**Table Appendices-6**) performed much better on the 2D plotting of the palato-alveolar fricatives /s/, /ʃ/ than the other map. In general '*ch-s-sh-i-u-o-lvq'* gave us better results; and this was the one used to start training the abnormal children as the therapist wanted also to focus on the /s/, /ʃ/ contrast.

V Permission from ethical committee and speech co-ordinator

It is required by law in Britain in order to test a certain drug or apply in general a new treatment method on a patient to show that there are not any potential hazards in the use of the procedure. OLT had to comply with such regulations. As such, a long procedure was carried on in order to get permission from an ethical committee to apply OLT. All relevant papers that were completed for the application to that committee can be found in **§Appendix-IX**. These include the ethics form that explains all the details about how we planned to apply OLT in speech therapy, the approval paper of the committee, as well as information sheets for the parents and the children.

The last stage before we started applying OLT was to recruit the clients and the therapist. This was arranged through Sheffield Speech and Language Therapy Services Paediatric co-ordinator. A special meeting was arranged for that purpose, a demo of OLTK was presented and a video playback of the normal child was shown to the speech co-ordinator. These were sufficient to get the final approval and the confirmation about the recruitment of the therapist and the clients.

VI Training of the therapist in OLT

Before the actual therapy started we made two rehearsals on the use of the application together with the therapist. We wanted to confirm that the therapist had grasped the main idea of how to use OLTK effectively and could carry on alone in the sessions with the client. In order to facilitate the training of the therapist in this new treatment method we prepared an on-line html version of the manual of OLTK that included a tutorial, demo section, and a troubleshooting guide. We have also written down instructions about the computer working environment that OLTK was running on, that in theory the therapist could start simply by turning on the computer. Finally we have given her lots of practical advise and suggestions for an effective use of OLTK.

6.2.3 The therapy program

Because of practical considerations the schedule and the description of the therapy stages in the ethics protocol was slightly modified. There were two main therapy stages and a total of nine sessions lasting one hour each. The third stage was the post-therapy review and appears in the next section with the results. The goal of the therapy was the successful realisation of /s/ at C, CV and CVC levels (C is the sibilant fricative and V is a vowel, **§Appendix-**XI). In teaching the correct gesture for /s/ the therapist gave instructions about the correct place of the tongue and she also took advantage of its similarity to /t/. Both are voiceless lingua-alveolar sounds. She also gave instructions about the position of the lips and the escape of the air from the incisors. The therapist was using OLTK for motivating the speech training, for receiving an additional evaluation and judgement on the articulation and for providing immediate feedback to the child by relating the gesture with 2D positions on the phonetic map. On the other hand the child was able to select his/her favourite animation game, “plane”, “dots” or "snaky line", and was trying to score as high as possible with the current setting of sensitivity by varying the articulation.

I First therapy stage

The first stage included a session with baseline recordings stored on a magnetic tape and digitally on the hard disk with the help of *SPEDATO* (§**Appendix-III-C)**. The list of utterances recorded (§**Appendix-XI**) was designed to sample the sibilant fricatives in a variety of phonetic contexts and some spontaneous conversation. Another session at that stage contained the aural tests where the therapist judged that the client had sufficient self-monitoring ability to distinguish the audible difference between a correct and an abnormal fricative production. The final session at that stage was devoted to demonstrating OLTK to the client and explaining the phonetic maps and the task to be accomplished .

II Second therapy stage

The second stage included five training sessions with OLTK lasting one hour each, twice a week over one month period (**§Appendix-VIII)**. This stage was the core of the whole therapy treatment with OLT. In this stage there was a mixture of computer-based and non-instrumental speech training.

A. Computer based training

During the previous demo session with OLTK the therapist identified the proper level of sensitivity to start each child. She found that at <–10> the child scored a 70% acceptance. This was adequate to start motivating the child to achieve better performance. The aim in all the following sessions was to encourage the child to make better and better productions at the same level of acceptance (70%) and correct classification of the speech patterns by raising the sensitivity. In order to achieve that, the therapist used the unlimited duration feature of OLTK to start a recording session with the child. Then she instructed the client on how to approach the target by giving aural and visual directions. Once the child had managed to reach the phone targets the therapist advised the child to sustain the production and repeat it. When the training on specific targets had finished, the therapist set a 20-30 sec recording time to test the new skills obtained. During that time the child had to remember and realise what was the improved articulatory position and repeat it as many times as possible in order to achieve a good score. A reward was also given by drawing stars on the bottom of the screen if the score exceeded a certain grade. If desired, the therapist could playback the recorded interval so that a better judgement could be made on the performance and the child could realise the mistakes or enjoy the success achieved. Each time an improved production occurred, the therapist saved it on the hard disk so that a record of all the successful efforts was maintained for further analysis and examination.

B. Traditional training

Despite the fact that OLTK kept children well motivated there were times where the child could not reach the target and the performance was poor. In these cases the child became easily bored and tired. Moreover we noticed that it was difficult for the child to concentrate on the computer environment and the instructions given by the therapist at the same time.

At those times it became necessary to draw the attention of the children and stimulate their interest with non-instrumental therapy methods. It is common in order to apply speech therapy to clients of that age to play together with them certain games. These games were based upon completion of a certain task by the child through successive rewarded efforts by the therapist. For example a game of TSENGA was adopted as suitable for that purpose so that for each removal of a tile the child had to say properly the sound or utterance in question, or an even simpler game with little plastic frogs, in which for each successful speech production, you were making a frog jump inside a bucket. One of the main differences here of course with the computer-based method was that the judgement was based solely upon the therapist and the feedback was not immediate. When the therapist had managed to make the child find the correct position of articulation, the training was continued again with OLTK.

This procedure was followed on average once or twice per session, with minimal time, 5-10min, and only if necessary. If the therapist judged that the child was improving and enjoying practise with OLTK, she was continuing the therapy with the computer.

6.2.4 Problems encountered

Many problems appeared during the therapy sessions with OLTK both technical and practical.

I The headset microphone

Perhaps the most significant practical problem was the fitting of the headset microphone. It proved to be rather inconvenient for the child, and sometimes the microphone arm was not at a proper position, which resulted in poor reception of the child's voice.

II The recording level

The level of recordings was also a very important practical and technical consideration. Although settings were provided to measure silence level, and a mixer tool could be used to set various levels on the sound card, these were not combined successfully. The noise from the computer's hardware, and also that from the hard disk during the recording buffering (§**4.2.1-I)** was certainly one of the sources of that problem.

III The loudness level

It was also the loudness level (§**6.2.2-IV**) that was causing OLTK to reject a spoken utterance; and often that could not be distinguished from rejection due to misarticulated utterances.

IV The accuracy of the phonetic maps

Finally we should mention here the problems encountered due to the specific design of the phonetic maps and the implementation of the sensitivity. As we have already mentioned above, at the start of training this sensitivity was set to a low value causing the system to accept pronunciations that deviated substantially from the target models. Problems started as the articulation of the child improved and the sensitivity was set to higher values. In that case OLTK could not discriminate successfully between the improved /s/ of the child and the norm of the model even for high values of the sensitivity level. In that point careful attention is needed by the therapist to realise that the production reached is still not the normal one, despite the fact that the target might have been reached with a high acceptance score and correct classification. As such no visual contrast or scoring distinction can be made any more between the two aforementioned sounds. Therefore a new map is needed with the ability to distinguish and create visual contrast between the child's new improved gesture and that of a normal sound. That led us to create individualised phonetic maps that included the improved skills for each child.

Individualised maps simply include a mixture of normal sound targets together with newly created sound targets from the improved efforts of the client. **Figure 6-1** shows an individualised map of ChildA that was used in therapy of the /s/ sound.

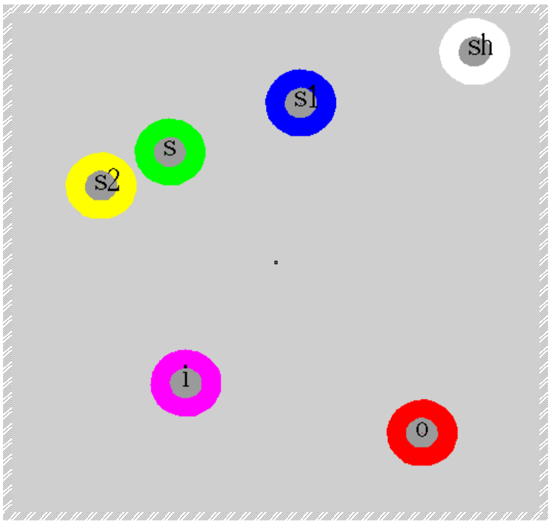


Figure 6-1 *An individualised map for therapy treatment of /s/ sound.*

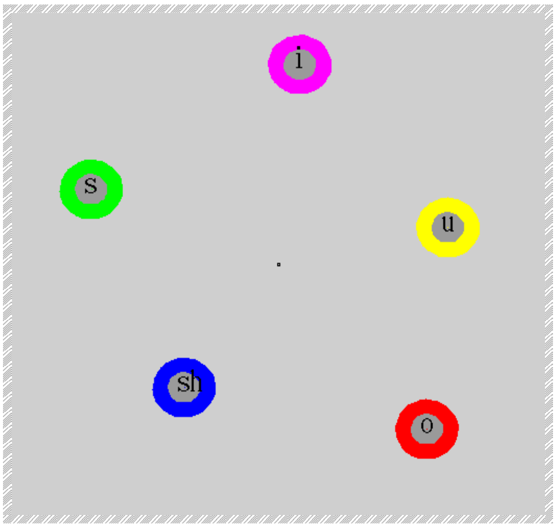


Figure 6-2 *Normal children's map for therapy treatment of /s/ sound.*

We can see the additional targets /s1/ and /s2/ in comparison with those included in the normal map **Figure 6-2**. These extra targets aim to provide sufficient visual contrast as different tongue positions occur for the palato-alveolar fricatives. /s1/ represents a palatalization of the /s/ sound: the air-stream escapes along a central groove in the tongue but the groove is not narrow enough, and the location is slightly to the back (case of recessive lisping). On the other hand sufficient lingua-alveolar stricture, which is the critical feature of normal /s/, occurs with the /s2/. In addition a secondary feature evident in that gesture is that the tip of the tongue is behind the lower incisors. The fricatives layout has been designed so that we can monitor and distinguish tongue retraction from front gestures like /s2/ to back gestures like /ʃ/. The same logic appears for the two vowels, front /i/ and back /o/.

In **Figure 6-3**, we can see the mapping of three abnormal productions of the /s/ sound on the normal children's map. In all three productions there is of course significant variation but the common thing is that target /s/ has been reached.

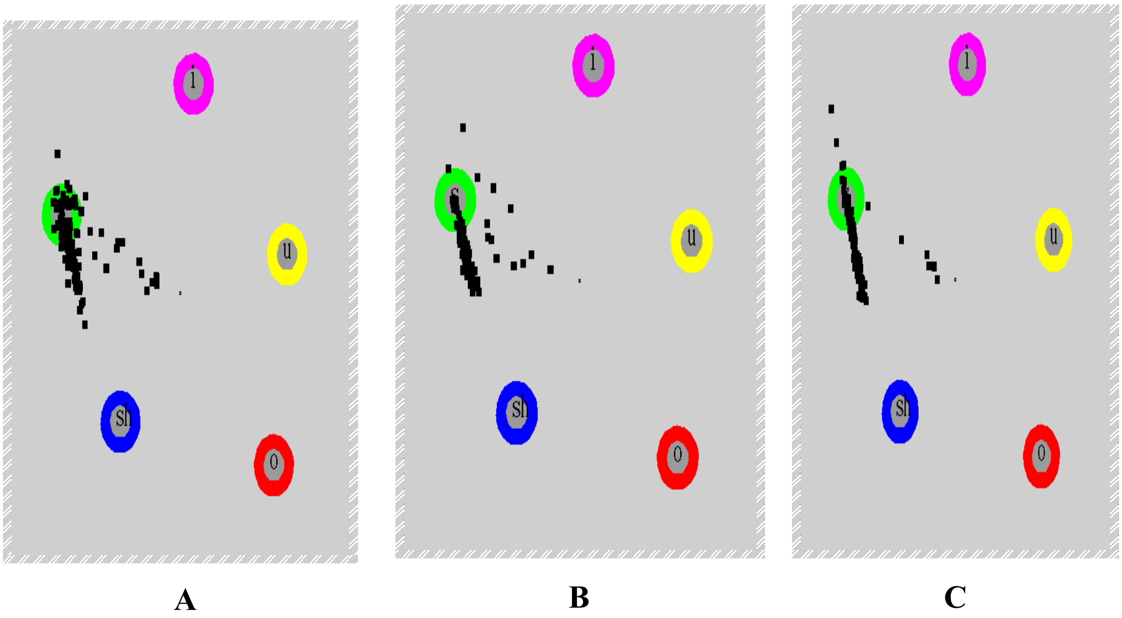


Figure 6-3 *Three abnormal productions of /s/ sound plotted on normal children's map. OLTK accepted the abnormal sounds and the area of the target models has been reached despite the high setting of the sensitivity level !*

A frame analysis of one of these recordings reveals the problem already described in the previous section. As we can clearly see in **Figure 6-4**, most of the frames are accepted with a sensitivity value above (-3). So although the sensitivity has been set to (-5) the current abnormal production scores higher than expected at a level where a normal /s/ sound is assumed. Clearly at that point the statistical model of the /s/ sound cannot distinguish between a normal production and an abnormal one neither visually or numerically. Therefore it is necessary to include the new sound in addition to those already comprising our map.

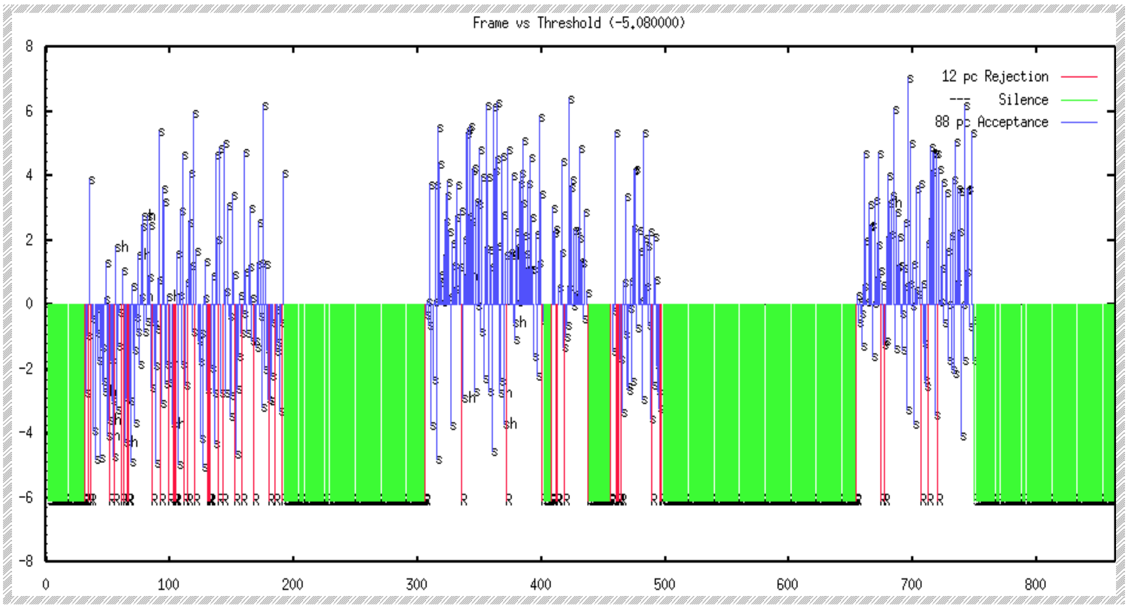


Figure 6-4 *A frame-by-frame analysis of an abnormal production of /s/ on normal children's map.*

The abnormal recordings shown in **Figure 6-3** are now plotted on the new map that is designed to create visual contrast and distinguish successfully between normal and abnormal production (**Figure 6-5)**. Variation is still visible but we can see that most of the utterances hit the blue target /s1/ that represents the abnormal production. Numerical differences and classification are also shown on the frame-by-frame analysis presented in **Figure 6-6**.

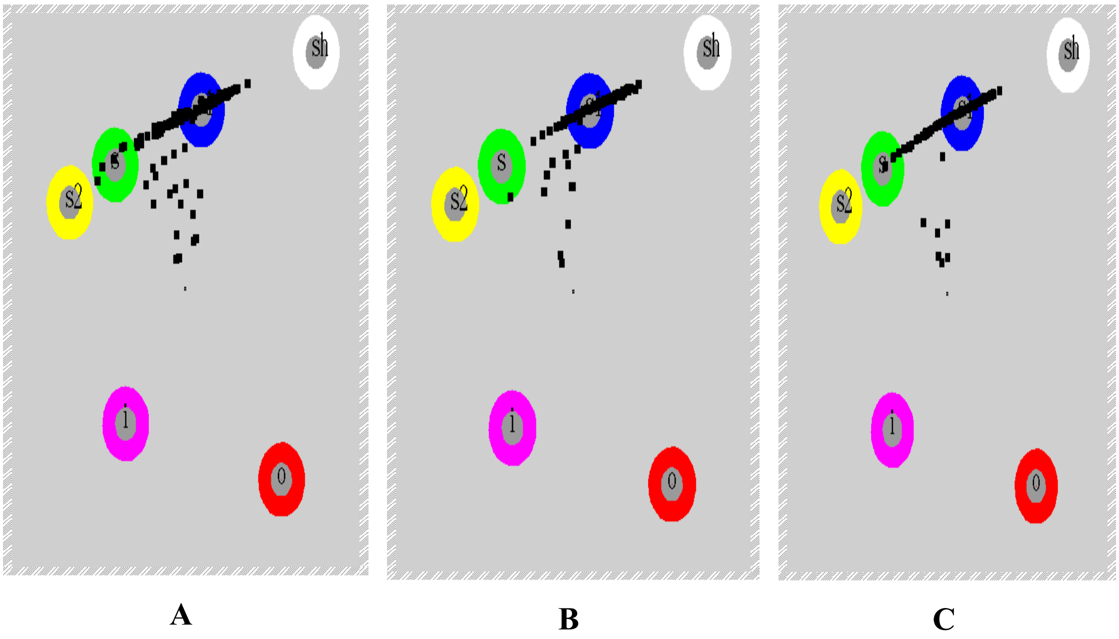


Figure 6-5 *A consistent type of /s/ abnormal productions is modelled so that OLTK can successfully contrast between normal and abnormal /s/ sound on a child's individualised map.*

The abnormal sound has been totally accepted but successfully classified to the new model of the map.

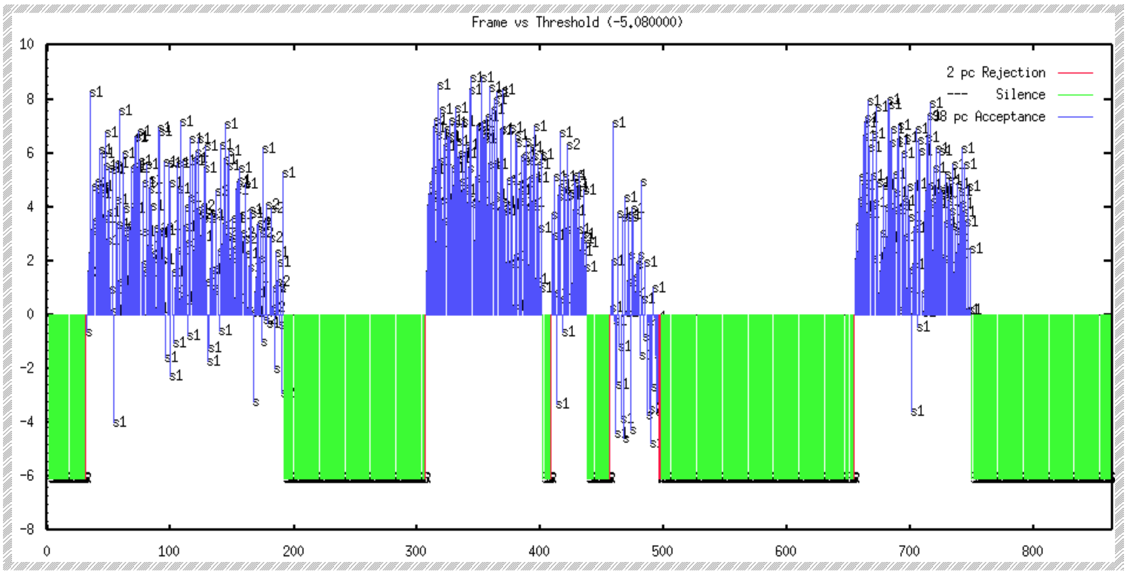


Figure 6-6 *A frame-by-frame analysis of an abnormal production of /s/ on child's individualised map.*

The evaluation of OLT is also corrected. As we can notice from **Table 6-1** on normal children's map the abnormal /s/ production have been mostly accepted and classified as a normal /s/, while on the individualised child's map with the inclusion of target /s1/ the same productions are again accepted but classified mostly as /s1/.

			Conditional Percentages					**Joint Percentages**				
Utterances	**Reject**	**Accept**	**/s/**	**/s2/**	**/s1/**	**/ʃ/**	**Check**	**/s/**	**/s2/**	**/s1/**	**/ʃ/**	**Check**
Figure 6-3-A	29%	71%	95%			5%	100%	67%			4%	71%
Figure 6-5-A	**8%**	**92%**	**2%**	**5%**	**93%**	**0%**	100%	**2%**	**4%**	**86%**	**0%**	92%
Figure 6-3-B	32%	68%	96%			4%	100%	65%			3%	68%
Figure 6-5-B	**6%**	**94%**	**4%**	**1%**	**95%**	**0%**	100%	**4%**	**1%**	**89%**	**0%**	94%
Figure 6-3-C	12%	88%	89%			11%	100%	78%			10%	88%
Figure 6-5-C	**2%**	**98%**	**8%**	**5%**	**87%**	**0%**	100%	**8%**	**5%**	**85%**	**0%**	98%

Table 6-1 *Evaluation and comparison of three different configurations of abnormal /s/ production by OLT on children's normal map and child's individualised map. The first two columns shows the percentage of frames rejected/accepted according to the HDF criterion - $P_{(HDF)}$. The next set of columns titled "Conditional Percentages" shows the posterior probabilities for the classification of the frame - $P(C_i|X)$. The last set of columns titled "Joint Percentages"is the product of the two events - $P_{(HDF)} * P(C_i|X)$.*

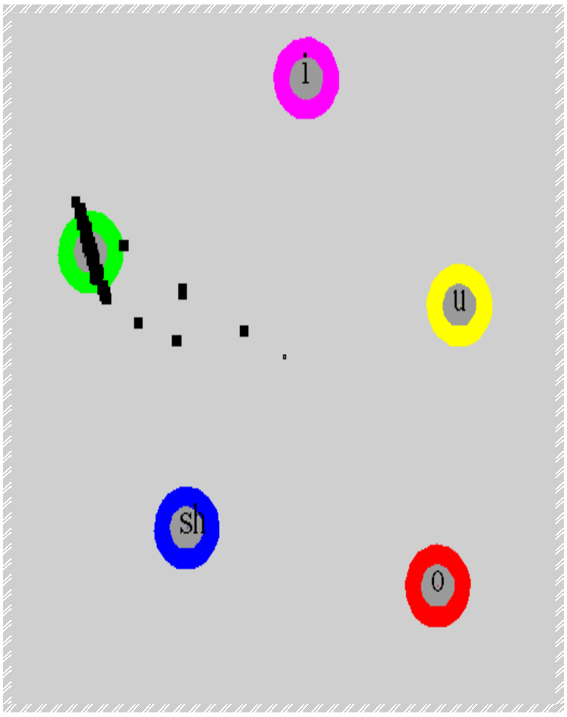


Figure 6-7 *Normal /s/ production on normal children's map with the blade of the tongue at alveolar ridge.*

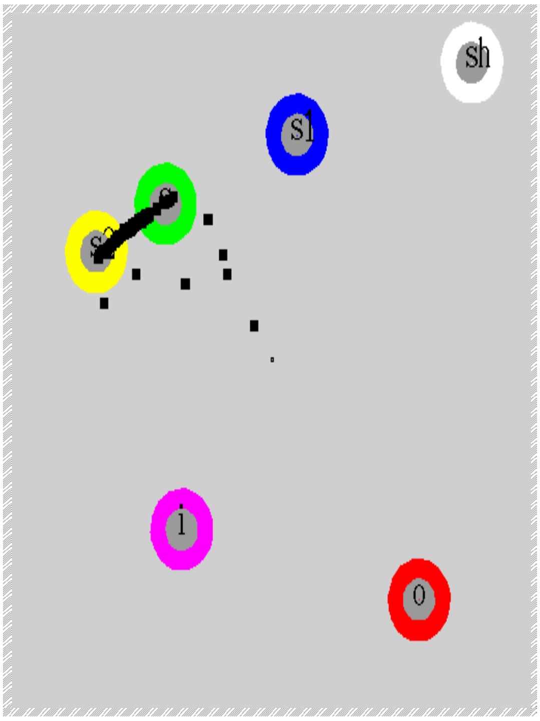


Figure 6-8 *Normal /s/ production on child's individualised map with the blade of the tongue at alveolar ridge.*

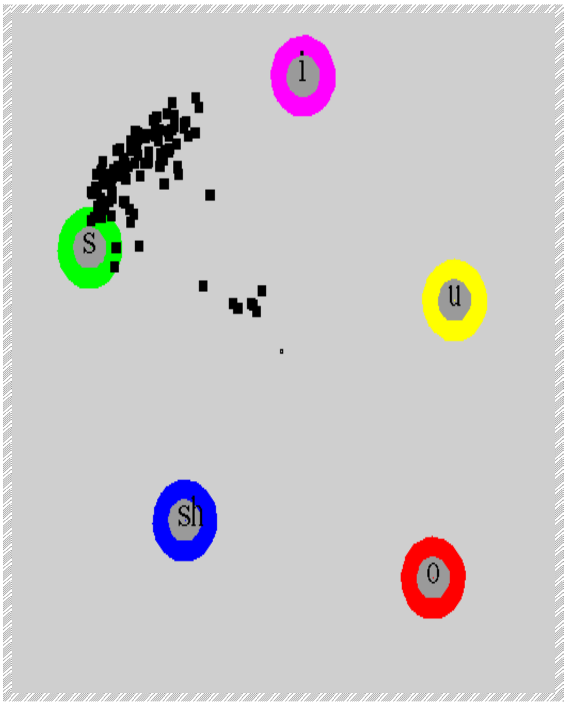


Figure 6-9 *Normal /s/ production on normal children's map with tip of the tongue behind lower incisors.*

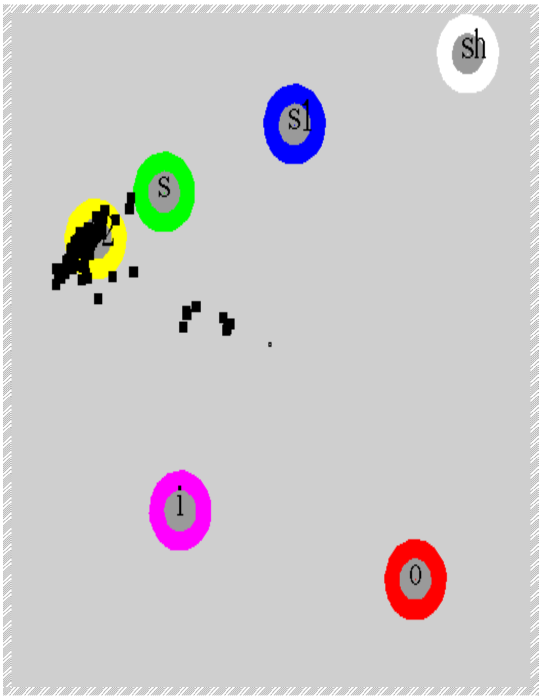


Figure 6-10 *Normal /s/ production on child's individualised map with tip of the tongue behind lower incisors.*

Not only can the construction of an individualised map distinguish between normal and abnormal gestures but it can also distinguish between different types of a normal gesture. For example see how we can compare visually two different types of normal /s/ production on both a normal children's map (**Figure 6-7, Figure 6-9**) and on a individualised map (**Figure 6-8, Figure 6-10**). It is obvious here that the statistical modelling of the /s/ sound from normal children reflects better the alveolar configuration for a normal production of that sound. The lower-incisors gesture is mapped quite unsuccesfully on the normal children's map (**Figure 6-9**): all the hits fall between the /s/ - /i/ area. But this inaccuracy was handled with the inclusion of /s2/ configuration on the targets of our individualised map (**Figure 6-10**).

Finally, as we can see from **Table 6-2**, OLTK evaluation of the child's utterances on the individualised map is improved. For example compare the first two rows of that table. The first row shows that the lower-incisors mapping of a normal /s/ gesture (**Figure 6-9**) is only 86% accepted and classified as 86% /s/ and 14% /i/. The second row shows that the same gesture under same sensitivity (**Figure 6-10**) is now 99% accepted and classified as 100% /s2/.

			Conditional Percentages					**Joint Percentages**				
Utterances	**Reject**	**Accept**	**/s/**	**/s2/**	**/s1/**	**/ʃ/**	**Check**	**/s/**	**/s2/**	**/s1/**	**/ʃ/**	**Check**
Figure 6-9	14%	86%	86%			14%-i	100%	74%			12%-i	86%
Figure 6-10	**1%**	**99%**	**0%**	**100%**	**0%**	**0%**	100%	**0%**	**99%**	**0%**	**0%**	99%
Figure 6-7	7%	93%	100%			0%	100%	93%			0%	93%
Figure 6-8	**1%**	**99%**	**2%**	**88%**	**10%**	**0%**	100%	**2%**	**87%**	**10%**	**0%**	99%

Table 6-2 *Evaluation and comparison of two different configurations of normal /s/ production by OLT on children's normal map and child's individualised map. The first two columns shows the percentage of frames rejected/accepted according to the HDF criterion - $P_{(HDF)}$. The next set of columns titled "Conditional Percentages" shows the posterior probabilities for the classification of the frame - $P(C_i|X)$. The last set of columns titled "Joint Percentages"is the product of the two events - $P_{(HDF)}* P(C_i|X)$.*

Individualised maps of children were created after the end of the third training session with OLTK, as we had by then collected a sufficient number of recorded utterances to retrain the maps and include models for more specific articulatory configurations. Careful examination and categorisation of the utterances for each one of the children recorded was a laborious task. From a recording session certain characteristic parts that relate to a new improved gesture had to be identified and then extracted and put into separate files to build up our new data for the newly created sound target on the map. For such a task, the visual analysis tools provided (§**Appendix II-C-b)** proved to be of great help: the unique features of OLTK for playing back parts of the utterance, even on a frame-by-frame basis, and also at the same time, the visual and aural examination, made a huge impact on our decision for the labelling of the new data.

Same type of individualised maps can be seen in **Figure Appendices-44**, **Figure Appendices-45**, **Figure Appendices-48**, and **Figure Appendices-49**. In all maps and results from ChildA and ChildB the first target named /s1/ corresponds to an improved configuration that these children managed to reach, and the second target /s2/ represents the successful efforts of the children to pronounce accurately the palato-alveolar fricative /s/. In the case of ChildC /s1/ reflects the same configuration as that of /s2/ for the other two children. Only one extra target was needed in that case as he developed his skills faster to reach sufficiently the normal /s/ sound of the map. Actually we could have created only one mixed target model from the speech data of both the client's attempts and the normal speakers of our map.

6.2.5 *Results*

I *Before and after therapy*

The results described in this section are based on the tables of **§Appendix-V**. Each table contains details of the utterances saved by the therapist for each one of the training sessions of a child with OLTK. The utterances were saved during each session, when the therapist judged that the performance was improving. All the tables describe the attempts of the clients for a successful realisation of /s/ in isolation. (There were also a few attempts of the clients at chaining /s/ with a vowel, but the number of recordings were not sufficient to present any data for analysis.) An average was calculated from all the recorded efforts after the end of each session for the acceptance/rejection levels as well as for all the fricative targets of the map. The graphs in the figures below shows these average percentages for each session for each one of these targets and for the combined good productions on both /s/ and /s2/ targets.

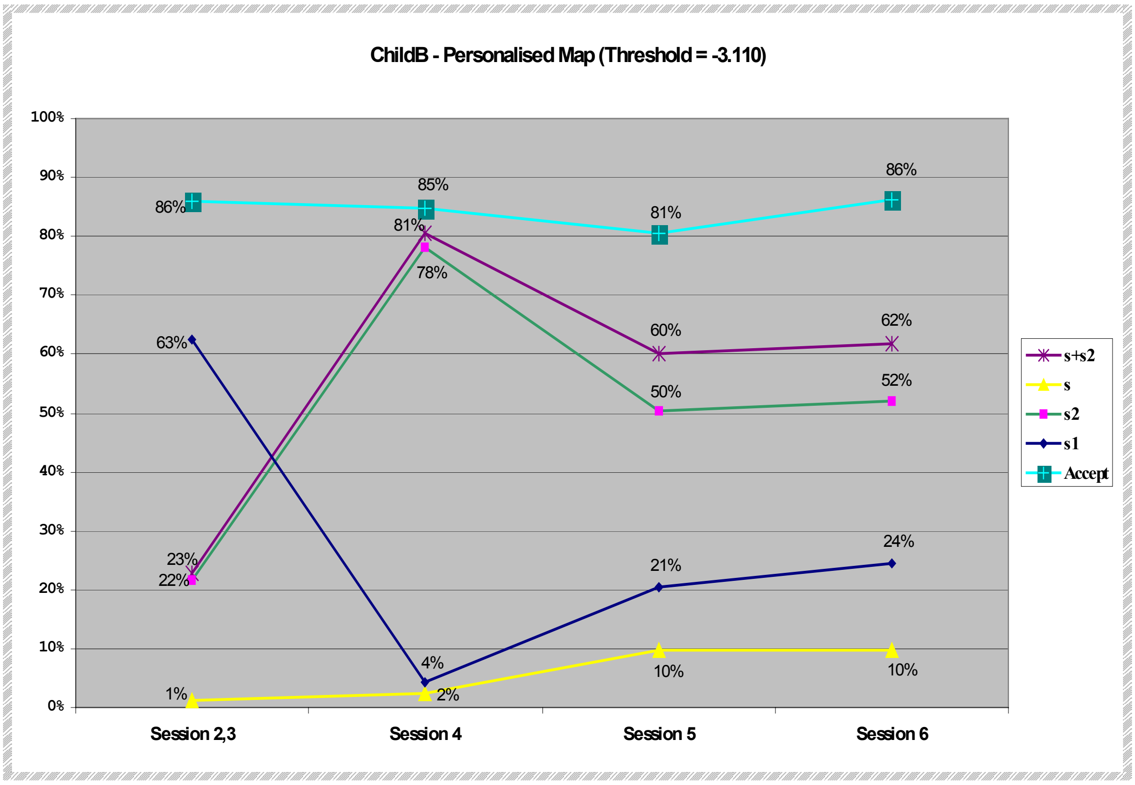


Figure 6-11 *ChildB average results per session for the fricative targets and acceptance.*

Figure 6-11 above describes the performance of ChildB on the *'ChildB_s1s2_lvq'* individual map (**Table Appendices-17**, **Figure Appendices-44**, **Figure Appendices-45**). The acceptance is maintained on almost the same level for all sessions and we expected it to be high for all the cases since we have included extra targets for abnormal production. From the other lines we can see that the bad productions marked with /s1/ have been considerably decreased from 63% to 24% while the good productions of both /s+s2/ have been increased from 23% to 62%. The increase in the performance is not monotonic: that means that the child could not realise successfully and maintain the best performance reached on the 4th session. This is discussed in more details in conclusions.

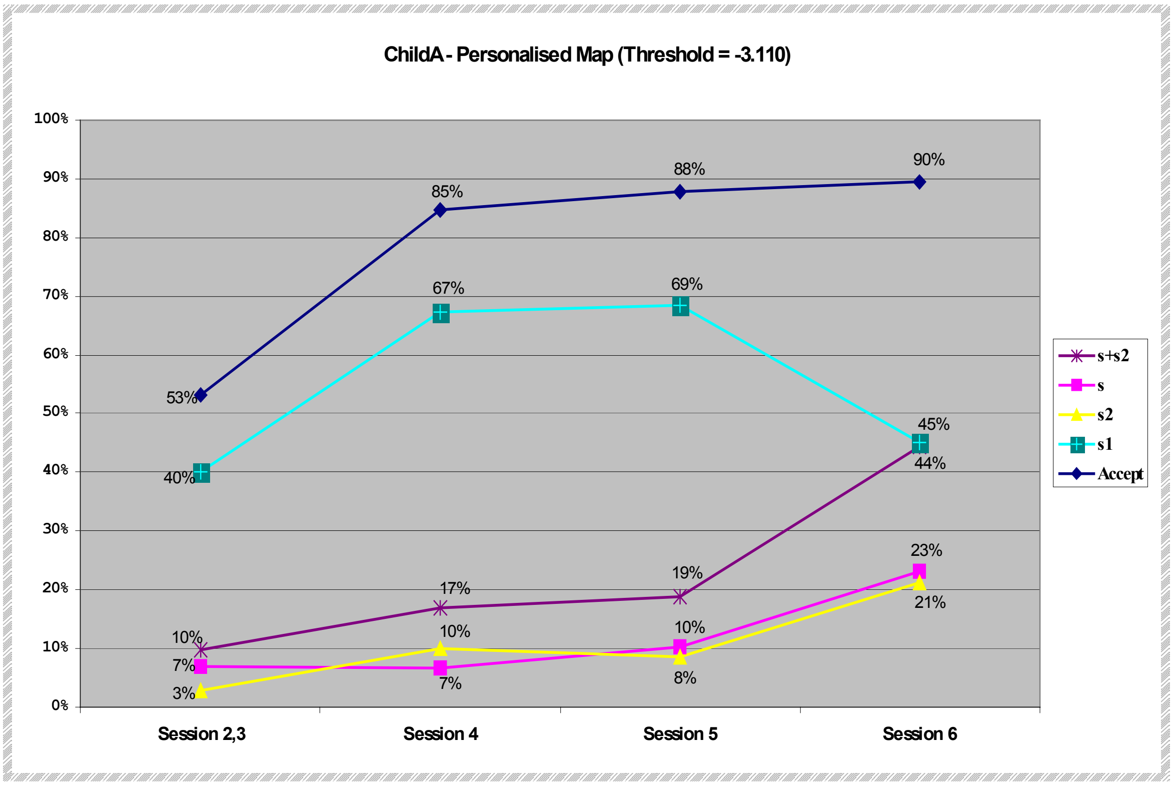


Figure 6-12 *ChildA average results per session for the fricative targets and acceptance.*

A monotonic increase in performance has been noticed in ChildA while testing his recordings on *'ChildA_s1s2_lvq'* individual map (**Table Appendices-8**, Figure Appendices-46, Figure Appendices-47). As we can see from **Figure 6-12**, /s/, /s2/, and /s+s2/, were gradually increasing reaching the maximum at the last session. Moreover, the acceptance on the first two sessions was relatively low, while the average of bad productions /s1/ increased in the 4th and 5th sessions but decreased in the last session to the same level as that of the 2nd and 3rd. A general observation here is that there has not been much improvement in comparison with that of the previous child. Some possible reasons are discussed in conclusions.

The third child, ChildC, was able to attend only three sessions; therefore there was not enough data to produce a similar graph for him. Despite that fact his progress was much faster and by the last session he was able to utter /s/ much more accurately than the other two children.

Finally **Figure 6-13** compares the performance of children on good productions between the initial and final therapy sessions. As we can see, in all the cases there has been performance increase both on the normal /s/ productions as well as the client's improved /s/, productions.

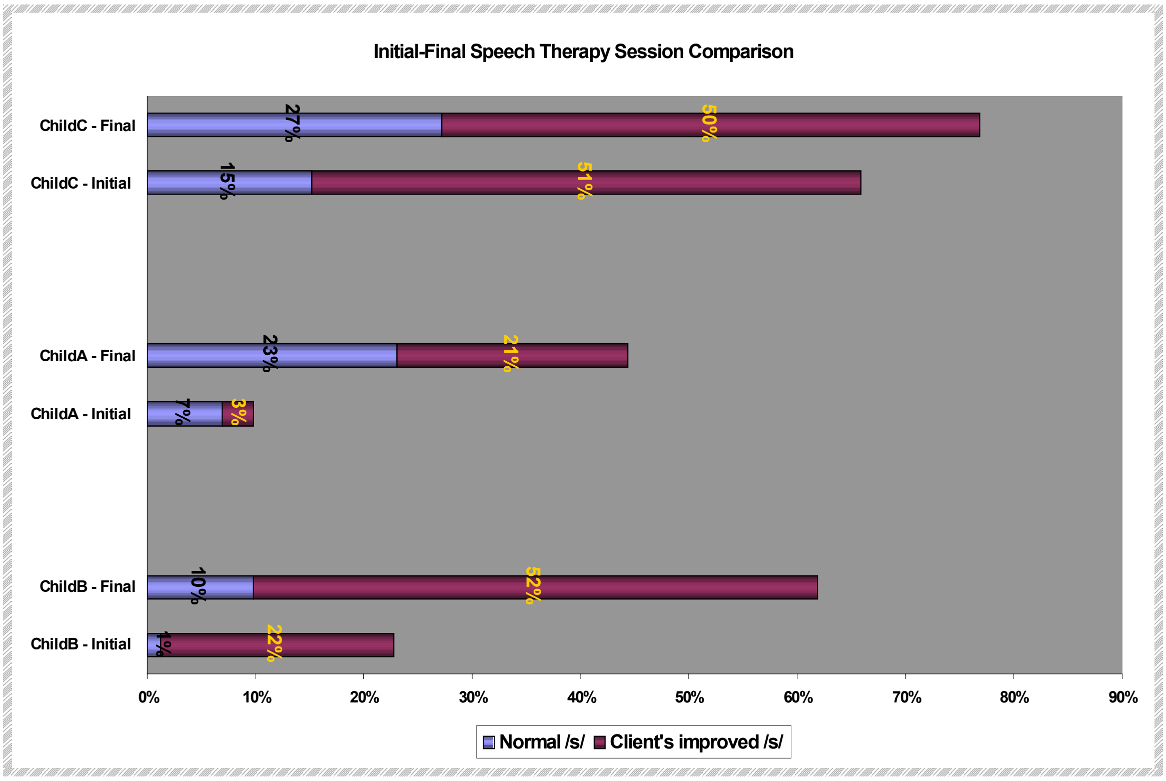


Figure 6-13 *Comparison of the good productions for all three children. The normal /s/ indicates what percentage of the children's attempts were judged similar to the /s/ target model of the normal speakers. The improved /s/ indicate the best productions reached during the therapy.*

II Post-therapy review : (Third therapy stage)

There were several tests to assess the speech production of the clients one month after the end of the speech therapy sessions. These included conversation, word pronunciation, and testing with OLTK. The therapist transcribed the sounds the children made and compared them with transcriptions taken at the date of the baseline recordings.

In two of the children, ChildB and ChildC there has been not only maintenance of the skills taught during the therapy session but also a significant improvement and consistency of proper articulation of the problematic configurations (**§Appendix-XI**). Parents reported that they were happy to see that gradually the speech of their children was improving. This was less evident on continuous speech where unconsciously the speech skills were difficult to maintain and predominate over the wrong skills that used to exist. But when careful thought and effort was put in to pronouncing a word properly the desired gesture was achieved with success. Testing with OLTK and the individualised maps of the children also showed great improvement as the children were able to reach the targets easily and achieve high scores with high sensitivity values around (-3) not only with isolated sounds but also with words. Another point to mention here is that the new skills acquired had a great effect to other similar gestures, like the voiced alveolar sibilant fricative /z/ where there had not been any significant time spent during the therapy sessions.

ChildA, showed little improvement from the original conditions before he started speech therapy. There are a couple of reasons to discuss here, one of which was the two different ways that he was instructed to articulate the /s/ sound. His natural tendency is to dentalise this sound and it seems the closer position to acquire a normal lingual-alveolar stricture would be with the tip of his tongue behind the lower incisors. Although he managed to do that, there was no consistency, and the therapist tried to teach him another way to acquire the normal /s/ gesture with the blade of the tongue at the alveolar ridge, a position similar to the /t/ sound. ChildA tried hard to achieve that, but unfortunately he retracted his tongue further back resulting in an abnormal palatalised /s/ sound. At the end of the therapy it was not certain which place of articulation would be more appropriate for him, but the individualised map created included targets to make him distinguish the different places of /s/ articulation.

6.3 Teaching accent modification to foreign speakers

The second application of OLT in speech training is to accent modification: particular the comparison of the English sibilant fricative production with that of a native Greek speaker and the process of modifying the second so to match the first.

We have to mention here that because of time limits the experiment has not been scheduled and designed in detail, and appears less formal than the previous study case. Nevertheless the results obtained were impressive for the few training sessions we had with the Greek natives.

6.3.1 Description of the Greek accent for sibilant fricatives

The description of the Greek gesture for the sibilant fricatives is more or less the same as that described in §**6.2.1-I**. The main difference between the Greek /s/ and the English /s/ is that in Greek the location of the lingual-alveolar stricture is relatively further back and the groove is wider than that of the normal English /s/ gesture, **[76]**. Concerning secondary features like the tip of the tongue, Petrounias states it is always behind the lower incisors but does not make contact with them, **[76]**. **Figure 6-14** shows a typical Greek accent of /s/ sound on the *'iso2vo4frlvq'* normal English male adult's map.

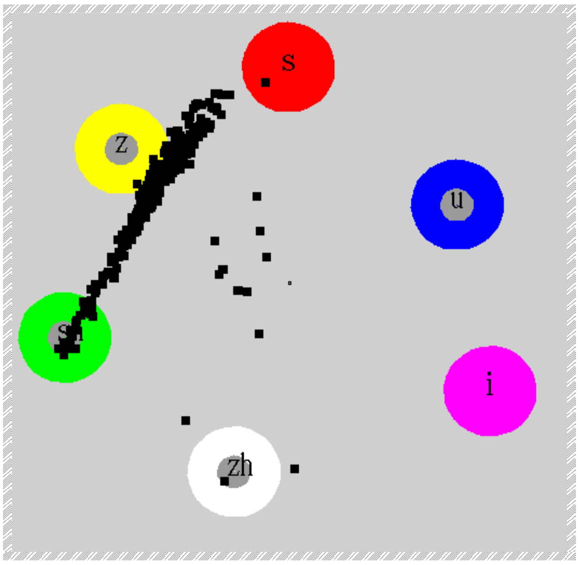


Figure 6-14 *Greek speaker's initial attempts to pronounce the English /s/ sound on the 'iso2vo4frlvq' map of normal English male adults.*

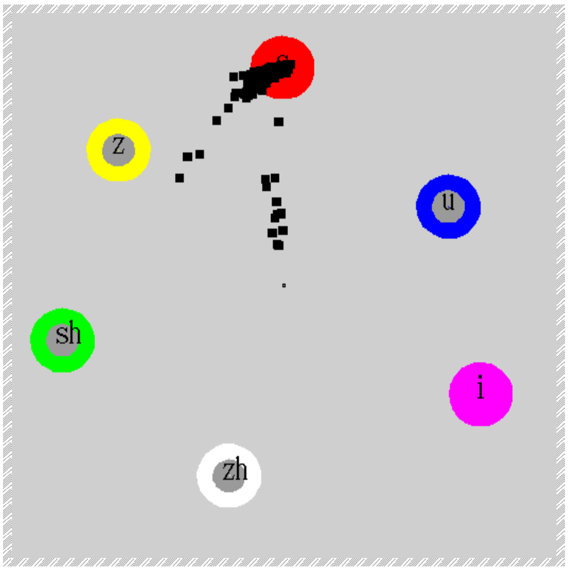


Figure 6-15 *An improved accent of the English /s/ sound by a Greek speaker on the 'iso2vo4frlvq' map of normal English male adults.*

The sensitivity has been set to (-20) level and only 53% of the recording has been accepted and classified as 46% /s/ and 54% /ʃ/. The 2D position of most of the hits on the map is somewhere between the two targets /s/ and /ʃ/ and that is quite consistent with the phonetic description we have given for the /s/ sound. We can also notice significant variation in the production of the sound, as many of the traces are scattered along the /s/ - /ʃ/ line.

6.3.2 Description of the foreign students

The four foreign students were all male adults, Greek native speakers with ages from thirty to sixty years old. Concerning their familiarity with the accent of the English language, AdultB and AdultC have lived in an English speaking community for one year and AdultD used to have casual, common English conversations because of his profession. Three of them AdultA, AdultB, and AdultC, have an English certificate. All the students had normal speech, hearing, vision and cognitive ability and no accompanying disabilities were noticed.

6.3.3 Target modelling

We used three maps: one with two fricatives /s/, /ʃ/ and four vowels, /i/, /u/, /a/, /o/, (**Figure Appendices-54**, **Figure Appendices-55**) another with four fricatives /s/, /ʃ/, /z/, /ʒ/ and two vowels /i/, /u/, (**Figure Appendices-50**, **Figure Appendices-51**) and another one with only two fricatives /s/, /ʃ/ and vowel /i/ (**Figure Appendices-56**, **Figure Appendices-57**). The layout of the maps was designed on the same principles as those in the therapy cases; therefore the targets on the map were modelled by recording the above sounds in sustained production. For that purpose we recruited eight normal English male adults. It is worth mentioning here that these speakers were all speech researchers and thus familiar with recording speech data and able to produce consistent configurations of the sounds in question. The recording format and method we used to record our data was the same with the one described in §**6.2.2-II**. Three datasets were used here as before (§**6.2.2-IV**) for training, testing of the map, and LVQ modelling of the targets. Accuracy of classification and mapping was also satisfactory. The numerical details for these maps can be found in the configuration tables (**Table Appendices-20, Table Appendices-22, Table Appendices-23**).

6.3.4 Independent speakers testing of the phonetic maps

Table Appendices-7 shows some tests we ran on the *'iso2vo4frlvq'* map. This one was most used in the sessions with our students. For that purpose we used two speakers, independent of the training of the map, and matching the physiological aspects of those recorded for the training, thus native English, male and adults. As can be seen from the table, the scores achieved for the attempts of the speakers to reach the normal targets, at the specific level of sensitivity, are high as expected. The exception was the productions of one of the speakers for the /z/, /ʒ/ sounds. Here we observed the following interesting phenomenon: the speaker was uttering the /ʒ/ sound perfectly but OLTK was classifying and mapping the sound as /i/ (**Figure 6-16, Figure 6-17**).

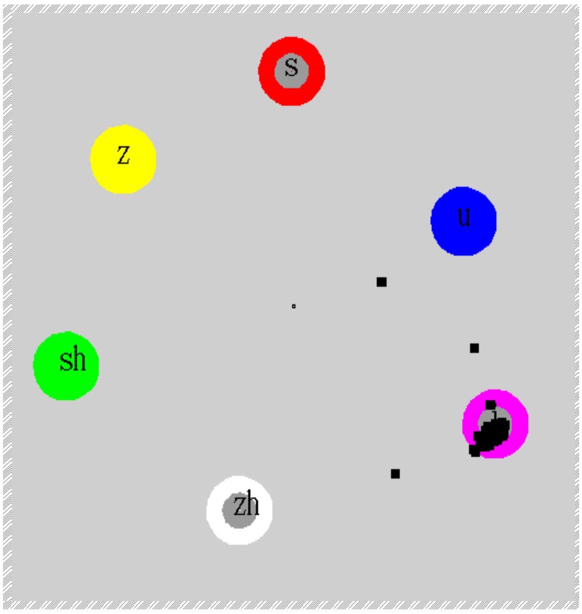


Figure 6-16 *Wrong mapping of a /ʒ/ sound.*

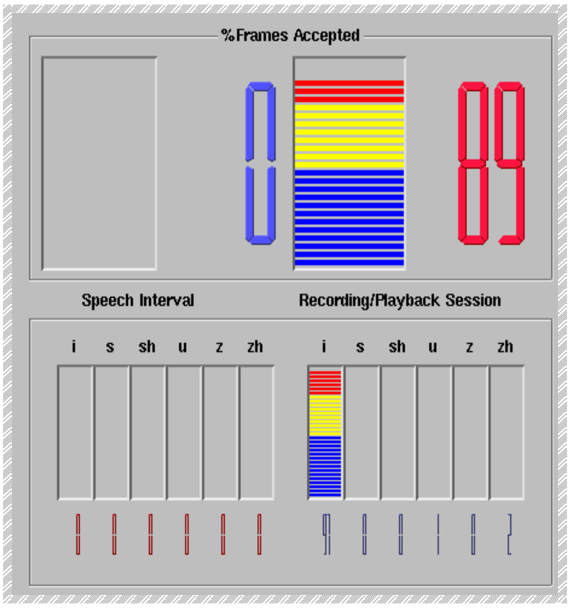


Figure 6-17 *Wrong classification of a /ʒ/ sound.*

The reader has to recall our discussion on the generalisation of the neural networks responsible for the mapping and classification **(§4.6.4)** as an explanation to the problematic behaviour of OLTK for the sound above. If we examine the two gestures for the /i/ and a /ʒ/ sound and if we try to simulate with our vocal organs how it is possible to glide from one sound to the other keeping the same labial setting, we will observe that the main feature that changes is the level of the articulatory stricture of the tongue with the palate. A /ʒ/ sound that is articulated with a relatively wide lingual-palate stricture can have a certain resemblance with a configuration of a high-front /i/ sound. In our case we trained the network only with certain typical configurations of /i/ and /ʒ/ sounds that certainly do not represent the extreme cases we have described. The new incoming unseen data presented passed our criterion for acceptance since they are quite similar to one of the modelled sounds, and then the net responsible made reasonable attempt to map and classify them. At that point the extra information appearing on our data, -extra voicing, wide lingual-palate stricture of /ʒ/- fools the net in to believing the new data belong to the /i/ cluster. So this results on the wrong mapping and wrong classification seen at **Figure 6-16** and **Figure 6-17**.

The quality of the map has also been assessed with a skilled phonetician in order to examine the ability of OLTK to reject sounds with a high sensitivity setting. In the same table (**Table Appendices-7**) we can see the classification and acceptance scores of utterances like the Greek /s/ sound, a palatalised or dentalised /s/ and a /th/. In all cases the map has successfully rejected the sound produced.

Last we made a final test with a Greek speaker to see if there is the potential to make him realise visually the audible differences between the two contrasting sets of sounds for English and Greek. The speaker with verbal explanations of the differences in production and with demonstration of OLTK soon became enthusiastic and managed to respond well in the various tasks we assigned with OLT.

6.3.5 Teaching objective

Awareness of the differences that appear on the English sibilant fricatives from their Greek counterparts as well as visual contrast at C and CV-CV context was the goal of the training. Self-monitoring and realisation of the articulation process that takes place during speech production creates the effect of learning the new skills which is required for uttering those sounds properly.

6.3.6 Description of the speech training sessions

The whole experiment took place within two weeks, and we arranged three individual sessions for all students, each one lasting approximately one and a half hours. In the first session we made base-line recordings; for that purpose we turned off the visual feedback of OLTK and recorded the various utterances for 30 secs each. In **Table Appendices-11**, **Table Appendices-12**, **Table Appendices-13**, **Table Appendices-14** the first attempt in the isolated sounds indicates a base-line recording.

The author acted both as instructor and demonstrator of how the OLTK software application works. He explained with verbal and non-verbal communication the production of English sibilant fricatives and made real-time recordings to show the students how to accomplish their task, which was to reach the targets on the map within a certain sensitivity limit. Previously recorded samples from native English speakers were also played back for further help and comparison. An important aspect was that, since the map in question could be used by both the students and the instructor, audio-visual comparison could be made between the two, especially when we used the ‘dots’ or the ‘snaky lines’ modules of OLTK. In this way the teacher could instruct the student to avoid certain gestures by audio-visually demonstrating to him the error, or to give him appropriate audio-visual feedback by demonstrating his correct position of articulators.

The first session ended by setting the sensitivity at (–20) and trying to experiment with unlimited time duration for the correct articulation of the sibilant fricatives in isolation. In the following two sessions our aim was to instruct the students with the method mentioned above so that they could improve their articulation by scoring above 70% for a higher sensitivity setting resulting of course in correct mapping and classification of the uttered sounds. The samples recorded in **Table Appendices-11**, **Table Appendices-12**, **Table Appendices-13**, **Table Appendices-14** for each one of the students reflect those effort as they have been saved when the students were managing to attain a higher score than that of their previous best attempts. In all of the samples recorded the duration has been set to 30 sec, and the student’s aim was to vary the articulation and repeat the utterance as many times as possible in that period so that the previous score achieved could be beaten. The second and third session included also CV-CV contrastive context utterances. The last setting of the sensitivity was at –5 level.

6.3.7 Results

Although as we have already mentioned that the time interval we devoted to training our students with OLT was small, the results obtained were quite impressive. To start with, **Figure 6-15** pictures the improved accent on the /s/ sound of a Greek speaker. As we can see, the target has been reached successfully and the variance in production comparatively with **Figure 6-14** is small. This time 84% of the utterance has been accepted and classified 100% as /s/ with the same sensitivity setting as that in **Figure 6-14**.

Analytical results for all the students trained and their performance can be found in **Table Appendices-11**, **Table Appendices-12**, **Table Appendices-13**, and

Table Appendices-14. All the corresponding graphs in **Figure Appendices-36**, **Figure Appendices-37**, **Figure Appendices-38**, and **Figure Appendices-39**, show significant improvement for all the speakers between the first and the last session of training in all the utterances recorded. From those graphs we can see that all speakers appeared to have problems in learning to pronounce correctly the /ʒ/ sound and this was apparent during the training sessions. In that case we found that it was helpful in improving their articulation for /ʒ/ if they started from a front, high, /i/ and gradually narrowed the stricture, ending up with the /ʒ/ sound. Trainees enjoyed also a lot the 'snaky' animation of OLT and this proved to be quite useful for the CV-CV utterances. An attempt to illustrate that animation with a series of pictures can be seen in **Figure 6-19**.

The average of the scoring for all the attempts of the students in the initial and final training gave us the graph of **Figure 6-18**. In general we can see that last session performance is at least twice as good as the first for all trainees.

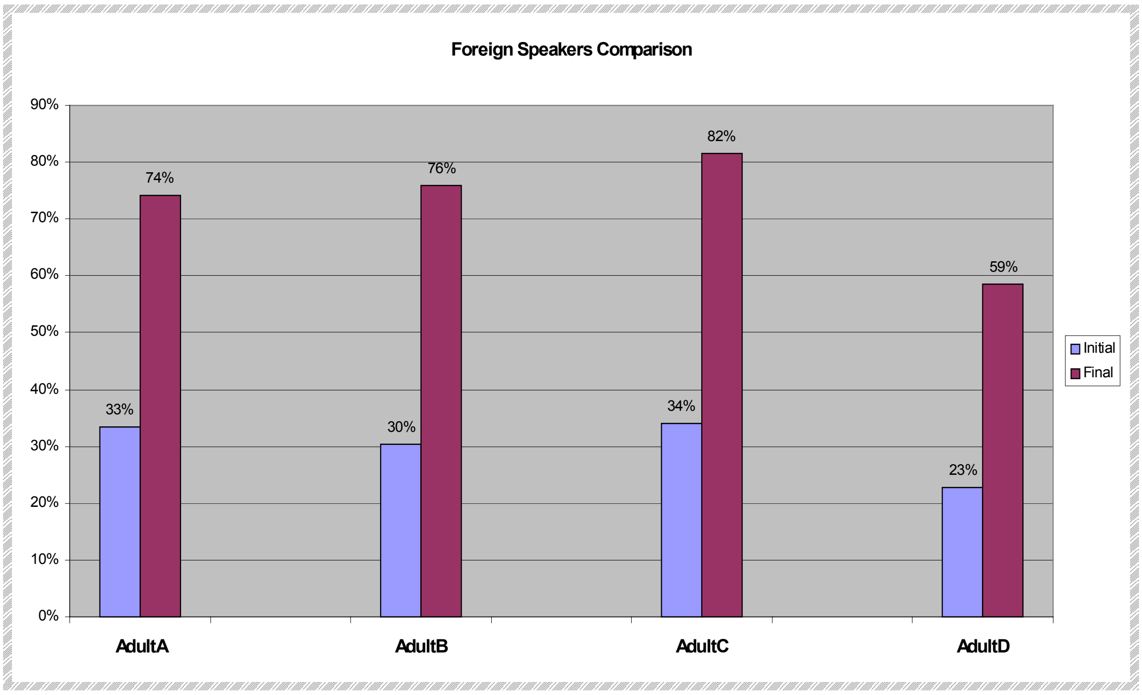


Figure 6-18 *Accent modification results of Greek native speakers on their initial and final training session.*

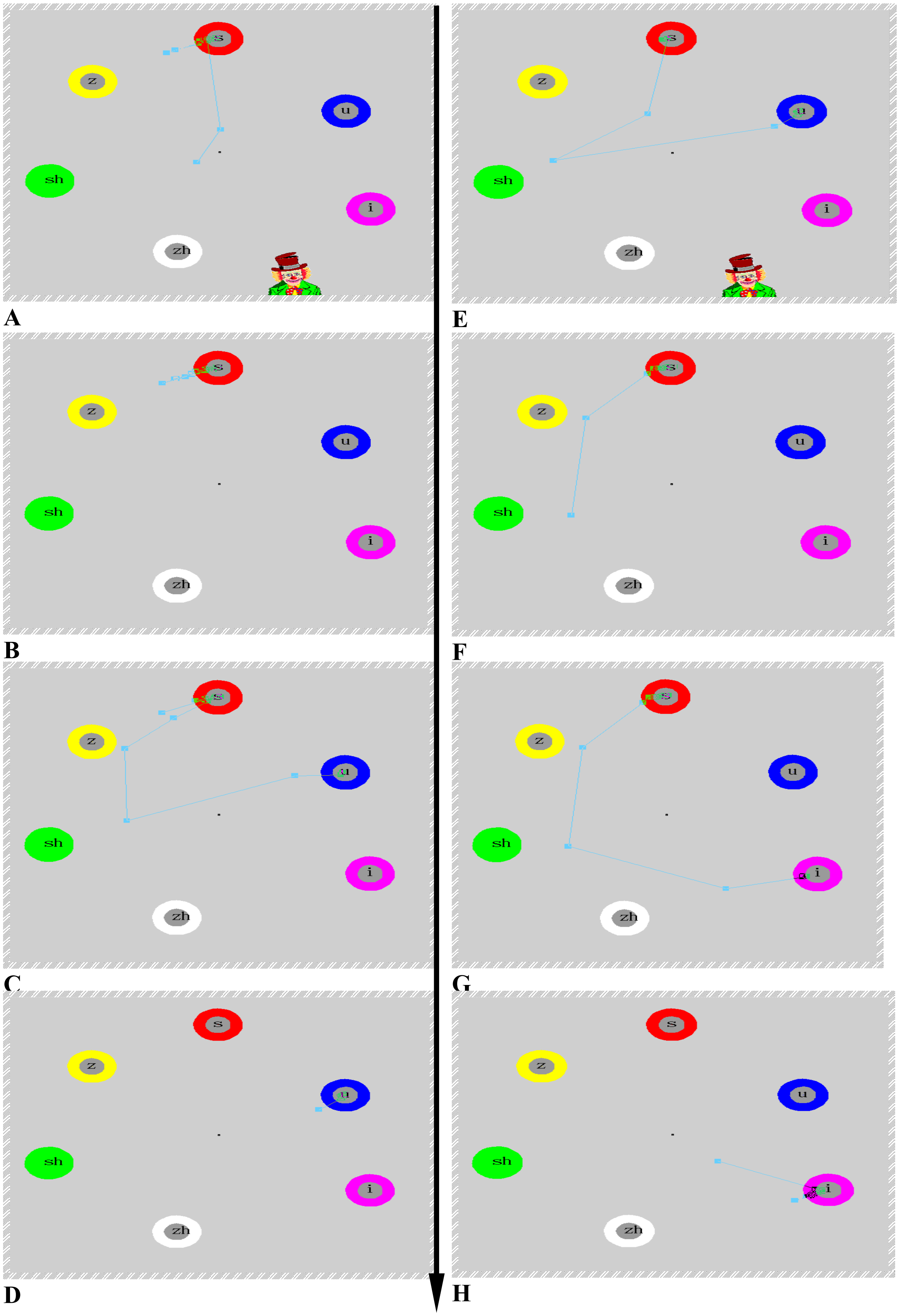


Figure 6-19 *Trajectories of utterances /su/ (Figures A-D) and /si/ (Figures E-H)*

Chapter 7

*Indeed, it would seem that he who disbelieves this bodily rising of the Lord is ignorant of the power of the **Logos** and Wisdom of God. If He took a body to Himself at all, and made it His own in pursuance of His purpose, as we have shown that He did, what was the Lord to do with it, and what was ultimately to become of that body upon which the **Logos** had descended? Mortal and offered to death on behalf of all as it was, it could not but die; indeed, it was for that very purpose that the Saviour had prepared it for Himself. But on the other hand it could not remain dead, because it had become the very temple of Life. It therefore died, as mortal, but lived again because of the Life within it; and its resurrection is made known through its works.*

St. Athanasius-On the incarnation (Translation-C.S.Lewis)

Conclusions

In the previous chapters we have discussed the Optical Logo-Therapy method (OLT), and the Optical Logo-Therapy toolkit (OLTK). Although the term OLT can imply all the theoretical aspects for a computer-based treatment model in speech training, we have used the term to refer mainly to the most critical feature of such a system which is the visual feedback provided. Likewise despite the fact OLTK can be considered as a computer-based speech training system specialised for articulation problems, the application of it in the areas of speech therapy and second language learning served the purpose of testing OLT visual feedback with real life problems. With this concept in mind, we proceed to analyse how successful we have been in our principal aims (§**1.10**) and what lessons we learned from the application of OLT in speech training.

7.1 The OLT feedback in speech training

The need for speech training emerges mainly from the inadequate self-monitoring abilities of the speaker to understand his/her errors in speech production and the inability to control his/her speech organs to correct these errors. The role of the instructor is first to understand the relationship that exists between the audible errors and the inaccurate control or position of the articulators. Then s/he has to provide two forms of feedback; navigational to direct the attempts of the client towards the correct place and manner of articulation, and evaluative to judge the outcome of each attempt and reinforce the new skills learned. The OLT instrumental method has been designed and implemented to fit in that activity basis, so that both the trainer and the trainee can benefit by the application of it. The theoretical framework of OLT embodies important treatment principles that have already been applied with success in various other instrumental and non-instrumental speech training programmes (§**2.3**, §**1.5**). We have discussed in general these principles in (§**1.6**) and now we are going to review how they are associated with OLT.

7.1.1 Visuomotor tracking

In OLT the client attempts to drive consistently the events on the visual display, e.g. fly a plane on a track, or blacken a specific area on the map. The visual patterns produced and the animation aim to reflect the immediate changes in the place and manner of articulation. Learning the correct utterance happens on a trial-and-error basis with: directions from the instructor (1), and correlation of OLT audio-visual feedback with the consistent control of articulatory features (2). While (1) can lead the client sufficiently close to the desired target production or enable him/her to understand how to reach the target, (2) can also fine-tune his/her attempts in order to gain precise control of the articulators and attain the exact configurations of an utterance.

7.1.2 Visual contrast

The *visuomotor tracking* principle of OLT becomes even more powerful when enhanced with contrast. Consistent error patterns and improved articulatory configurations can be contrasted with the normal targets on the two dimensional phonetic map of OLTK by including extra targets to tailor and personalise OLTK to the performance of the individual. This way the teacher can advise the client to avoid 'hitting' or passing over certain areas on the map to eliminate the misarticulated patterns, and encourage the production of utterances that produce the desired visual events. It is important to emphasise that this visualisation of speech events resolves the ambiguous relationship between auditory perception to the audible differences of speech production. Vision becomes the primary modality for comparing utterances since acoustic similarities are converted to the equivalent topological proximities. It is both the two dimensional distance from the sound target for contrast purposes and the proprioceptive cues associated with the direction of the movement on the 2D phonetic map that accomplish the navigational feedback.

7.1.3 Reinforcement

The effect of OLT navigational feedback is to increase the frequency of correct and accurate responses. The drills are built in such a way that they fulfil a computer game oriented strategy where the acquired speech skills depend on how skilfully the user drives the visual events on the display. This approach which we have characterised as qualitative (§**2.2**-{D}) is coupled with the definition of a metric for a quantitative evaluation of the speech quality and a built-in connectionist network for identifying segments of speech. The evaluative feedback reinforces even more the attempts of the client as s/he tries to beat the previous score record of his/her best production or best recording session. In that case, a reward triggers the interest to the maximum and compensates for the hard efforts of the client.

7.2 Evaluation of OLT visual feedback

To answer the important question concerning the effectiveness of OLT visual feedback, (EVF), we have to check whether OLT complies or not with the requirements we set down in §**2.2**.

7.2.1 OLT real-time audio-visual feedback

We have analysed in depth the software mechanisms behind the real-time processing cycle of OLT (§**4.8.2-III**). We concluded that both the recording and the playback process of OLT can work in real-time providing that the resources of our computer system are such that can bear the load of OLTK software. This requirement, {A}, has always been in the top list of our priorities. While it is common in non-instrumental speech training, and it may be the case in other computer-based speech training systems, that the feedback is provided *after the productions have occurred*, in OLT it happens *as they occur*. This has the interesting property that the client establishes cognitively an immediate association between the acoustics of speech production, the visual events from the mapping transformation of speech events, and the articulatory mechanisms. In particular, the navigational feedback of OLT is such that varying slightly the articulators has as a consequence the small movements on the phonetic map that correspond to small changes in the acoustics of the speech produced. This temporal equivalence of speech perception, speech visualisation, and speech production happens almost instantly, in milliseconds. Therefore we understand now how important it is to capture these micro changes in the spectrum of speech in order to be able to analyse the acoustics and visualise the speech events concurrently. It is OLT real-time processing and visual feedback that makes it easier for the client to realise the errors during speech production by looking at the animation/drawing on an OLT map than listening to the misarticulated sounds.

7.2.2 *OLT speech features*

It is requirement {B}, that raises the issue of what particular speech characteristics to isolate and focus on. OLTK uses cepstral analysis (§**4.4**) a well-known modelling of the time-varying spectral envelope of speech. Since we are interested in articulation, this method provides us with a number of features that aim to represent differences in the acoustics produced by the time-varying properties of the articulators. As such the amount of detail extracted from the speech signal and represented on the map can be neither too low, resulting in a very specific speech training that does not cover all aspects of articulation, nor too high level resulting in confusion and incomprehensible visual representations. However cepstral analysis does not take into account nasality and lip rounding. These speech characteristics play often a very important role and cannot be ignored. Moreover place and manner of articulators based on this kind of speech representation can be inferred only implicitly from the acoustics of the sound produced and not explicitly as for example in an articulatory synthesis model. Of course in such a computer-based system, extracting more information about the articulators solely from the acoustic input of microphone is a difficult task. We will explain more about it in the following subsections.

7.2.3 *OLT visual representation*

There are two reasons for the significance of visual representation, {C}. First it should be natural for the client to make a logic and consistent link between changes in his/her articulation and changes in the display, and second the visual representation has to be designed to be as simple as possible, therefore comprehensible. OLT visual representation is based on phonetic modelling. Certain sounds that appear to be problematic in the phonetic inventory of the client are modelled by collecting samples from normal speakers and then mapped on predefined fixed positions on two dimensions as simple geometric shapes, (circles, ellipses, boxes). Although in practice this representation appears to be attractive and simple for the clients to understand what the target is that they have to reach, in theory there are a lot of problems. The most important is the absence of a clear and direct relationship of the articulatory gestures and the correlated phonetic area defined and displayed on the map. In other words, speech training focus on minimal distinctive sounds, not minimal distinctive articulatory features. This is an old-fashioned treatment method that is gradually abandoned in favour of the modern theoretical approaches that study the features of articulation and the related phonological processes, (§**6.1**). Indeed, even the definition of what is a phoneme or what is a phone is obscure. What is considered to be the norm of a phoneme in a language? How much it can be varied without changing its identity? What is the relationship between abnormal sounds, or any other sound that the human can produce with the phonemes of a specific language? All these questions makes us realise that it is not convenient to teach sounds to the students, rather it would be much better if we teach how to produce these sounds, how to vary the articulators in order to achieve a sequence of well defined gestures. In theory OLT has the potential for such a representation and we have seen that even with the limitations of the current phonetic modelling it is possible to visualise certain features, like lingual stricture, lip rounding, or tongue retraction, with the accompanied inconsistencies of course.

7.2.4 OLT navigational feedback

Requirement {D} is strongly affected by {C}. It states that there should be visual indications, and directions shown on the speech display showing how to correct the error and reach the target. This is one of the weaknesses of most computer-based speech training systems and OLTK is not an exception. Nevertheless with the help of the instructor and the individualised maps, the students can gradually move away from the targets of misarticulated patterns to the normal ones. Currently the space on the map between two targets is not well defined. As we have discussed in §**4.6.4**, the net responsible for the mapping does not have enough information to generalise for the map area between the phone clusters. Consequently systematic variation of specific articulatory features may result in systematic but not consistent drawing/animation on the phonetic map due to the inadequate modelling and visual representation.

7.2.5 OLT contrasting models

The power of OLT is mainly in the visual contrast, {E}, between the mapped utterance of the client and the target models defined. These models are composed of the normal productions of other speakers or the improved utterances of the client. However the contrast is between broad sound categories, phonemes or phones, and it is not defined explicitly for articulatory features. That means also that it is difficult to create visual contrast for syllables, treating them as a combination of more than one phonemes, since the area between any two phonemes in our map is not well defined. The cause for that is the modelling of phonemes from sustained speech productions that do not take on account how one sound glides to the other, (§**6.1**).

7.2.6 OLT evaluative feedback

In OLT the evaluation, {F}, of each attempt of the client to utter correctly the target sound or the syllable depends on how similar the incoming sound patterns are to the target models, how well they 'fit' into our subset of speech data space. Clearly we can draw the conclusion that the creation of the data space and the influence it plays on the calculation of such a metric is substantial (**§4.5.1 - §4.5.5**). We should also make clear that in the present version of OLTK the *goodness-of-fit* metric is not associated with the 2D distance of the sound mapped from the 2D centroids of the sound clusters. Another feature of OLT evaluation is the classification of the speech input and the visual transformation of this information as heights of the vowel-consonant bargraph display **(§3.1.2-II-B)**. The heights show the possible confusion of the incoming sound with the model sounds of the map. It is possible to train the client with only that kind of display **(§Appendix-II-D-b)** but it has not been utilised in our experiments.

7.2.7 OLT motivational impact

Both the drawing/animation on the phonetic map, and the evaluation provided with the bar displays together with the reward, motivate sufficiently the student to attain the desired response, {G}. The design of the display as a video-game and the various options for altering the appearance of the targets and the animation/drawing hold the attention and the interest of the clients (§**7.3.2**).

7.2.8 OLT flexibility and suitability

The phonetic map can be tailored to include those sound targets that the student get confused and others that represent his/her improved state, {H}. Once the map is tailored and the sensitivity of OLTK is set on an appropriate level the student is able to reach target production without discouragement.

7.2.9 OLT accuracy

The accuracy, {I}, of OLT visual feedback is strongly affected by two intermingled factors, the modelling of the sound targets and the level of *sensitivity*. These issues are discussed analytically in §**4.5.6-II**. In brief, for very low settings of the sensitivity there is the danger that the student's deviant accent will become easily accepted but mapped inconsistently; or, on the other hand, to become often rejected resulting possibly in the student's disappointment and frustration. Moreover, the sound input may be acoustically too similar to the sound models of the map, yet phonetically considered to be a different sound. In such a case OLTK may not be able to distinguish it from the sound models represented on the map no matter what is the level of sensitivity. From all these problems the one potentially most hazardous in the speech training is the one where the feedback provided is faulty, relating the acoustics of a production with wrong positions or areas of the map.

7.3 Evaluation of OLTK in speech training

Despite the limitations and the problems about OLT we have discussed, a piece of software that provides OLT feedback, OLTK, has been applied successfully to the cases of functional articulatory disorders (FAD) and accent modification (AM). Before we review the results of the application and the issues related to OLTK we would like to refer briefly to the objective and the features of the software.

7.3.1 The objective of OLTK

What we consider to be normal speech productions are achieved through the realisation of the relationship that exists between the different articulatory gestures and the acoustics of a certain speech production. OLTK can help the user to obtain such a realisation and awareness through the visual and oral modalities. More specifically, it aims to create visual contrast between different articulatory gestures and to provide visual monitoring of how one can go from a certain articulatory configuration to a different one by varying certain articulatory features.

The key points to consider in applying OLT feedback are first the use of real-time phonetic displays to make trainees to become aware of the contrast between different sound configurations and second to reveal to an extent the place and manner of articulation so that they can refine their articulation with certain movements of the articulators.

In its present state OLTK can be considered as a very specialised computer-based speech training system that has the following main features.

1. Real-time audio and visual animated feedback during recording and playback, synchronised with on-line evaluation of the client's articulation.
2. Simultaneous phoneme and speech segment training. Sound awareness and phoneme comparison and contrast with models from other normal speakers and client's improved utterances.
3. Visual comparison with previously recorded utterances.
4. Visual monitoring of the variation of articulatory features.
5. Visual perception of different gestures through the animation/drawing on different areas of the phonetic map.
6. Motivation and reinforcement of clients through the use of graphics and game like strategies with several animation types.
7. Specialised phonetic maps tailored according to the physiology of the client and the speech disorder.
8. Interactive control of various parameters governing the visual feedback.

9. Frame-by-frame analysis of articulation enhanced with special graphics displays and combined with traditional ones like speech waveforms and spectrograms.

7.3.2 Review of speech therapy with OLTK

The application of OLT in the field of speech therapy can be considered the most difficult and the most demanding in terms of the peculiarities of speech training. Our clients were children and that increased all the precautions we took to guarantee the smooth running of the software and its functionality. Part of this involved the training of the therapist on the new method to apply for treatment. Despite our efforts spent to make the software as user-friendly as possible, the complexity of the computer environment and the rather technical instructions we provided initially the therapist had a rather limited effect to her understanding concerning the use of OLTK (**§Appendix-XI**). Nevertheless practice makes perfect, so at the end she could use most of the options of the software and run the therapy sessions without any help.

During the first two or three sessions OLTK was mainly used to maintain motivation and build confidence. The therapist reports (**§Appendix-XI**) that features of OLTK such as animation, reward stars, evaluation score, and threshold level, *"focussed the children on maintaining high scores and trying to beat previous threshold levels"*.

As the therapy was progressing due to the intensive reinforcement of OLTK the children quickly established increasingly accurate consistent articulatory patterns (**§Appendix-XI**). At the same time problems such as loudness, position of the microphone and recording level became more obvious (§**6.2.4**) but the most critical one, namely the accuracy of OLT, could hardly be noticed. The therapist complained that; OLTK loaded with the general map of normal speakers *"failed to reject instances of lateral productions of /s/ despite a high threshold level setting and awarded PM high scores"*. This is one of the cases we referred in **§7.2.9**. The danger was well spotted (**§Appendix-XI**); *"without reference to perceptual analysis reliance on OLT feedback would have reinforced incorrect placement of the articulators…"*. Unless we had built an individual map including the lateral sound in our map configuration the therapy would have stopped (§**6.2.4-IV**). The creation of individual maps increased the accuracy of feedback, reflected the progress made and according to the therapist, *"were an essential aid to the process of fine-tuning and establishing consistency of production"*.

Although many of the main features of OLTK have been used in the speech therapy, others have not been fully utilised (§**7.3.1-4, 5**). We think this occurred because at the time of the experiments, we were not in position to explain adequately to the therapist how the different areas of the map can be associated with different gestures and how it is possible by altering articulatory placement to glide from one sound target to another. Hence the therapist did not encourage enough the client to observe what effect would have on the map if s/he was experimenting by altering the place and manner of articulation.

Because of this we are left with the impression that during the therapy the children were focussing more on the score produced and the reward given. *"They tended to get very engrossed in winning reward points and to enthusiastically pursue the same placement"* (§**Appendix-XI**) rather than looking at the map and trying to pursue different placement of the articulators.

Even with the approach followed the results at the end of all the therapy sessions showed a successful realisation of /s/ in isolation for all the children, (§**6.2.5-I**). The results obtained from the measurements with OLTK about the improvement agree with the transcriptions of the therapist (§**Appendix-XI**) and the reports from the parents of the children (§**Appendix-X**). These results were really very encouraging concerning the progress of the children but most important was the fact that two of the children maintained the gains and showed that they have even increased them at a post-therapy review, (§**6.2.5-II**, §**Appendix-XI**). Success in generalising to situations outside of the treatment sessions is typically described as a gain of 'carry-over' of articulatory learning. Achieving carry-over is perhaps the most difficult stage in articulation remediation, **[45]**, and the efficacy study with the children showed that OLT measured improvement there too.

The final remark of the therapist concerning the use of OLTK was that she found it an extremely useful tool in the treatment of misarticulated fricatives. The parents of the children had the same opinion concerning the feedback their children were receiving (§**Appendix-X**). In addition all parents replied that their children were interested in the drills with OLT and greatly enjoyed the treatment (§**Appendix-X**).

7.3.3 Review of accent modification with OLTK

The application of OLT in the area of second language learning, has also been successful, (§**6.3.7**). Because of the maturity of our students we had the opportunity to make even more intensive speech training and experiment a lot more by instructing them on how to vary their articulation. Soon the students established a good relationship between the visual patterns appearing on the map and the acoustics of their speech production. This time the clients focused more on the map and less on the score displays and they obtained the new skills required for accent modification without the need of building individualised maps.

7.3.4 OLTK autonomy

Although the issue of autonomy is an important one in the design of any speech training aid, it remained out of the scope of this research as our primary aim was the theoretical and practical study of effective computer-based audio-visual feedback. Nevertheless the question of self-teaching is highly dependent on the quality and the accuracy of the feedback provided. As such in the present state unsupervised use of OLTK can be hazardous because of the problems we have reported in §**7.2.9**.

Perhaps we can be less strict on the autonomy of the system in accent modification problems. Actually this can be a good test before we decide its use in the area of speech therapy. Alternatively, once the therapist is certain that a specific set-up (sensitivity levels, map) of OLTK is not going to evaluate and present wrong feedback on the attempts of the client s/he can assign well described drills for home practice. Indeed, the report of the therapist confirmed that it may be possible to use OLTK in home maintenance activities once the software is further refined to require less specialist technical expertise (§**Appendix-XI**). The parents of the children also expressed their interest and reacted positively for homework with OLTK (§**Appendix-X**).

The benefit from unsupervised speech training is double, both because the therapist has time to see more patients and because the client can keep practising at home and stabilise performance before the next training session.

7.3.5 Other issues

Other issues like cost of the equipment and courseware are also very important and contribute significantly to both the autonomy and effective use of a computer-based speech training aid. OLTK can be considered to meet the requirements for the first but not the second. Moreover OLTK is also missing administrative modules necessary for storing and updating student records. Finally, OLTK is specialised in articulation, but a complete speech training program requires almost all aspects of speech, such as pitch, voicing, laryngeal quality, nasality. A detailed consideration of articulation, should verify that OLTK can be used with other classes of sounds, and extent the use of them in words, rather than only sustained and in syllables.

7.4 Future inspirations

OLT is still in the infancy. Our research showed both the weaknesses and the advantages of applying OLT in speech training. Given that there is a great prospective by introducing certain modifications and improvements in the method and the software, we can be assured that OLT has a bright future. Some of the things we plan to solve existing problems and enhance the functionality of OLT are:

7.4.1 Map design

Currently OLT depends upon the design of phonetic maps. We believe that for the reasons we reported in §**7.2.3** the notion of phone and phoneme has to be significantly diminished from the oncoming design of our map. Consider for example the vowel phonemes, *"part of the problem in describing vowels is that there are no distinct boundaries between one type of vowel and another"*, **[58]**. Therefore it is necessary to define, wherever possible, what are supposed to be the standard gestures, the norm gestures, for the phonemes of a language and also the range for the allowed variation of articulatory features that cause the variable forms of that gesture. Ladefoged observes that *"the phonetic characteristics of a sound cannot be determined by measuring the absolute values of the physical phenomena involved. Instead, we must state the percentage values of the range that is possible for each speaker"*, **[58]**. Hence the range should be defined as a continuous scale independently of the speaker.

In the actual implementation of the new phonetic maps a gesture is marked at a position on a predefined two dimensional fixed point. Then we use a specific 2D direction to map the allowed variation of this gesture according to the changes of articulators. A skilful phonetician can be our model in that case. Once the 2D area around the gesture is defined we can proceed by defining the area between any two sufficiently dissimilar gestures. This is the harder part, as we need a very precise method to capture all the intermediate configurations between the initial and final gesture. An idea is to try and incorporate some articulatory synthesis or simulation for creating our model data. Another is to find quasi steady state segments of speech and interpolate between the two. Visual symbols for interpreting the different areas on the map and special graphics to indicate how one can glide from one gesture to another will dramatically help the teacher to give appropriate directions and instruction to the student. From the client's point of view all these details can be hidden and the map can be shown more or less with the previous convenient abstract phonetic form and the graphics associated with. Hopefully this mapping will reflect better the change of certain contrasting features. The new technique for map design will be called acoustic articulography, (AA).

7.4.2 Consistent mapping

Certainly the modelling above will improve the consistency of mapping, but we expect those sounds out of the training data space will still cause problems if they are accepted. Therefore the individualised maps will keep playing an important role in the future versions of OLT. Of course the whole process for creating any map, collection, segmentation, labelling, speech processing, and training of the artificial neural networks for mapping and classification, should be automated in such a way that will not require any technical expertise.

7.4.3 Correct evaluation

Consequently we can hope for a more accurate evaluation once we have the ability of constructing individualised maps with the AA technique. This is because the 'goodness of fit' metric is based on the modelling of our data. Actually the map transformation will become even more meaningful if there is an analogy between the distances in the sound space and distances in the mapped space, something like multi-dimensional scaling. For, if this is combined with a visual effect of zoom in, zoom out, scaling the axis of the map will result in a fascinating way of visualising speech in two dimensions. Two dimensional evaluation will be much easier to calculate in this way and the qualitative navigational feedback will be more closely related to the deviant articulation.

7.4.4 OLTK improvements

There are really many, but we will discuss mainly about those that mostly influence the OLT feedback.

A. Loudness

If the overall energy of the signal is included in our speech parameters we have to make sure that this is not going to affect the articulatory evaluation of the client. In practice that means that there should be visual indications for 'too loud' or 'too soft' productions. There should be also easy handling of the recording level and the silence detection module of OLTK.

B. Computer hardware noise

The CSL (Computerised Speech Lab) hardware and Visi-Pitch II from Kay Elemetrics address the problem of "noise" from the computer by including an external input/output module to isolate the input signals from the noisy computer environment. We believe that such a facility should be incorporated in our system too.

C. Speech analysis and parametrisation

As we have explained in §**4.4** the type of analysis and parametrisation we followed is a standard method in many speech recognition systems. Nevertheless this approach should be argued rather than assumed since the speech training may require a different perspective as it concerns the modelling of the sound targets (§**6.1**). In particular it may require a different level of window and frame analysis depending on the sound in question, e.g. in plosives the frame period should be less to capture the closure and release phase. The therapist also may want the client to focus on syllables or phonemes or transitional parts of an utterance and thus vary the analysis in order to increase the level of detail in the visual information of the system. In conclusion a dynamic representation of speech features (§**7.2.2**) may be proven to be more beneficial in that case.

D. Sophisticated reinforcement and reward system

This should resemble more or less a game approach. The idea is stronger reinforcement and more compensating reward for the different level of treatment and client's performance. The instructor should have the option to set a threshold for the reward and define when and how is going to be awarded to the student.

E. Interface appearance

There are different groups of professionals and non-professionals who can be considered as the users of OLTK. These can be distinguished according to their involvement in speech training and what they are aiming for with the use of the software application. Typical categories of users are :

Instructors	Aim	Clients
Phoneticians	Teach articulatory phonetics.	*Students*
Language instructors	Teach accent modification.	*Foreign speakers*
Speech therapists	Treat segmental articulatory disorders.	*Patients*

Although speech training is common in all cases, the method and the approach can be quite different. Therefore we believe that OLTK interface appearance has to be differentiated and become more distinct for the various groups of people operating the program. Usually in other software applications there is an option where the user of the program can select the level of options detail.

Our close collaboration and advice from speech therapists, phoneticians, statisticians, and programmers was essential. We believe that for maximum creativity and innovation, future efforts to develop an effective computer-based visual feedback treatment and the corresponding speech training aid require a level of co-operation to be maintained between all these experts.

7.5 Epilogue

A quick look on the principle aims we set out in the beginning of this thesis (§**1.10**) and a comparison with the review of OLT in §**7.2**, reveals that OLT is indeed a real-time audio-visual feedback based solely on acoustic input (§**1.10-a**). Undoubtedly also, the motivational and visual reinforcement characteristics of OLT (§**1.10-d**) have been beneficial for those involved in speech training, instructors and students. The rest of the aims were partially met because of the problems that appeared. Certainly our intervention to create the individualised maps, their acceptance from both the therapist and the clients, and the successful outcome of speech training with them have been the factors that overcame these problems up to a limit. In particular the children could refine their articulation and reach the sound targets with instructions from the therapist and the visual contrast on the map (§**1.10-b**). There has been evaluation from the system (§**1.10-e**) on the attempts of the clients to encourage them to produce correctly the target sound and assist the therapist on her judgement. Because of the mapping limitations speech training focused more on sustained sound production, and the effect of visualising the variation of certain articulatory features (§**1.10-c**) has not been fully utilised.

To conclude, OLT can be considered effective in speech training; the important theoretical principles embodied, and the successful results from the therapy sessions and the accent modification case were encouraging and generally supportive of the idea that potentially OLTK may become the favourite and inseparable companion of the speech training instructor.

Bibliography

[1] Alvarez A., Martinez R., Gomez P., and Dominguez J. L. (1998), *"A Signal Processing Technique for Speech Visualisation"*, ESCA – Still 98, Marholmen, Sweden, May 24-27

[2] Anderson S., and Kewley-Port D. (1995), *"Evaluation of Speech Recognisers for Speech Training Applications"*, IEEE transactions on speech and audio processing, Vol. 3, No. 4, July.

[3] Arends N. (1993), *"The visual speech apparatus. An aid for the speech training"*, Manuscript, Instituut voor Doven, Nijmegen, pp. 135-156.

[4] Arends N., Povel D.J., Michielsen S., Claassen J., Feiter I. (1991), *"An Evaluation of the Visual Speech Apparatus (VSA)"*, Speech Communication (1991), 10, pp. 405 – 414

[5] Baer T., Gore J.C., Gracco L.C., and Nye P.W. (1991), *"Analysis of vocal tract shape and dimensions using magnetic resonance imaging: vowels"*, J.Acoust.Soc.Am., vol. 90, no. 2, pp. 674-679, 1990

[6] Ball M.J., Grone B. (1997), *"Imaging techniques"*, In Ball M.J., and Code C. (Eds), Instrumental Clinical Phonetics, Whurr Publishers

[7] Benedetto M.D., Destombes F., Merialdo B., and Tubach J.P. (1982), *"Phonetic recognition to assist lipreading for deaf children"*, Proc. IEEE ICASSP-82

[8] Bernstein J. (1989), *"Application of speech recognition technology in rehabilitation"*, in Speech today and tomorrow: proceedings of a conference at Gallaudet University, September 1988, (Ed). B.M. Virvan, pp. 181-187, Gallaudet University, Washington.

[9] Bernstein L.E., Ferguson J.B.III., Goldstein M.H.Jr. (1986), *"Speech training devices for profoundly deaf children",* Proc. ICASSP '86, TOKYO, pp. 633-636

[10] Bishop M., (1999), *"Latent variable models"*, in Learning in Graphical Models, Jordan M. (Ed.), NATO Science Series, MIT Press, Boston, pp. 372-402

[11] Blache, S. (1989), *"A distinctive feature approach"*, In N. Creaghead, P. Newman, & W.Secord (Eds.), Assessment and remediation of articulatory and phonological disorders (pp. 361-382). Colomubus, OH: Charles E. Merrill.

[12] Bloodstein O. (1984), *"Speech Pathology : An introduction"*, 2nd edition by Houghton Mifflin Company, Boston, MA

[13] Brooks S., Fallside F., Gulian E., and Hinds P. (1981), *"Teaching vowel articulation with the computer vowel trainer: Methodology and results"*, British Journal of Audiology, vol. 15, pp. 151-163.

[14] Browman C. P., Goldstein L., Saltzman E.L., and Smith C. (1986), *"GEST: a computational model for speech production using dynamically defined articulatory gestures"*. JASA, 80.

[15] Clark, J. and Yallop, C. (1991), *"An Introduction to Phonetics and Phonology"*, Oxford, Blackwell Inc.

[16] Dagenais PA, Critz-Crosby P, and Adams JB. (1994). "*Defining and remediating persistent lateral lisps in children using electropalatography: preliminary findings.*", American Journal of Speech-Language Pathology, Sept: pp 67-76

[17] Davis S.M., and Drichta C.E. (1980), *"Biofeedback: Theory and application to speech pathology"*. In N.Lass (Ed). Speech and Language Advances in Basic Research and Practice., New York: Academic Press.

[18] Duda R.O., Hart P.E., and Stork D.G. (1999), *"Pattern Classification (2nd ed.)"*, Ricoh California Research Center, California, USA.

[19] Ellis E.M., Robinson A.J. (1992), *"Two Dimensional Representation of Phonemes of the English Language"*, Proc. (IOA) Conference on Speech and Hearing 14(2), pp 407-414.

[20] Ellis E.M., Robinson A.J. (1993), *"A Phonetic Tactile Speech Listening System."*, Cambridge University Engineering Department CUED/F-INFENG/TR122, May.

[21] Emerick L.L, and Hatten J.T. (1989), *"Diagnosis and evaluation in speech pathology"*, Prentice Hall, Englewood Cliffs, NJ

[22] Fanty M., and Alleva F. (1993), *"Speech Tools User Manual"*, Center for Spoken Language Understanding, Oregon Graduate Institute of Science & Technology, Beaverton, Oregon, USA.

[23] Farmer A. (1997), *"Spectrography"*, In Ball M.J., and Code C. (Eds), Instrumental Clinical Phonetics, Whurr Publishers.

[24] Fausett L. (1994), *"Fundamentals of Neural Networks : Architectures, Algorithms and Applications"*, pp 289-300, Prentice Hall, London, U.K

[25] Fitzgerald M., Gruenwald A., Stoker R., (1989), *"Software review – Video Voice Speech Training System"*, Volta Review, vol. 89, pp. 171-173

[26] Fletcher S.G, Dagenais P. A, and Critz-Crosby P. (1991), *"Teaching Consonants to Profoundly Hearing-Impaired Speakers Using Palatometry"*, Journal of Speech and Hearing Research, Volume 34, pp 929-942.

[27] Fletcher S.G, Dagenais P. A, and Critz-Crosby P. (1991), *"Teaching Vowels to Profoundly Hearing-Impaired Speakers Using Glossometry",* Journal of Speech and Hearing Research, Volume 34, pp 943-956.

[28] Fowler C.A., Rubin P., Remez R.E., and Turvey M.T. (1980), *"Implications for speech production of a skilled theory of action".* In B. Butterworth (Ed.), Language Production I London: Academic Press.

[29] Gibbon F., and Hardcastle W. (1987), "*Articulatory description and treatment of 'lateral /s/' using electropalatography: a case study. British Journal of Disorders of Communication.*", Vol 22, pp 203-217

[30] Gold B., and Morgan N. (1999), *"Speech and Audio Signal Processing : Processing and Perception of Speech and Music"*, John Wiley & Sons

[31] Gomez P., Rodellar V., Alvarez A., Bobadilla J., Bernal J., Nieto V., Perez M. (1995), *"Estimation of speech formant-dynamics using neural networks"*, Eurospeech'95, Madrid, Spain, September 18-23.

[32] Grunwell, P. (1995), *"Clinical Phonology"*, London, Chapman & Hall, pp. 130

[33] Halle, M. and Stevens, K.N. (1979), *"Some reflections on the theoretical basis of phonetics",* in Lindblom, B. & Ohman, S. (Eds), Frontiers of speech communication research, Academic Press, London.

[34] Hardcastle W. J, Gibbon F. E, and Jones W. (1991), *"Visual display of tongue-palate contact : Electropalatography in the assessment and remediation of speech disorders"*, Brit. J. Disorders of Communication, 26, pp. 41-74

[35] Hardcastle W.J., and Gibbon F. (1997), "*Electropalatography and its Clinical Applications*" , In Ball M.J., and Code C. (Eds), Instrumental Clinical Phonetics, Whurr Publishers, 1997.

[36] Hatzis A., (1995), *"Visualisation of the articulation for the hearing impaired with self-organising maps (VAHISOM)"*, Unpublished MSc Thesis, The University of Edinburgh, Dept. of Artificial Intelligence, September.

[37] Hatzis A., Green P.D., and Howard S. (1996), *"Optical Logo-Therapy - (OLT) : A Computer-Based Speech Training System for the Visualisation of Articulation Using Connectionist Techniques."*, Proc. Institute of Acoustics (IOA) 18, 9, pp. 299-306.

[38] Hatzis A., Green P.D., and Howard S. (1997), *"Optical Logo-Therapy - (OLT) : A Computer-Based Real-time Visual Feedback Application for Speech Training"*, Proc. EUROSPEECH'97, Vol 4, pp. 1763-1766.

[39] Hatzis A., Green P.D., and Howard S. (1999), *"Visual Displays in Practical Auditory Phonetics Teaching"*, Phonetics Teaching & Learning Conference (PTLC'99), University College of London, U.K.

[40] Hawkins S. (1992), *"An introduction to task dynamics"*. In G.J. Docherty and D.R. Ladd (Eds) Papers in Laboratory Phonology II: Gesture, Segment, Prosody. Cambridge: CUP.

[41] Hegde M.N., and Davis D. (1995), *"Clinical methods and practicum in speech-language pathology"*, San Diego : Singular Publishing Group.

[42] Henderson, E.J.A (1971), "*Structural organisation of language: phonology"*, in Minnis, N. (Ed.), Linguistics at large, Paladin, London.

[43] Hewlett N. (1985), *"Phonological versus phonetic disorders: some suggested modifications to the current use of the distinction"*, British Journal of Disorders of Communication 20, pp 155-164

[44] Hewlett N., and Shockey L. (1992), *"On types of coarticulation"*. In G.J. Docherty and D.R. Ladd (Eds) Papers in Laboratory Phonology II: Gesture, Segment, Prosody. Cambridge: CUP.

[45] Hoffman P.R., Schuckers G.H. (1983), *"Articulation remediation treatment models"*, in Articulation Assessment & Treatment Issues, Daniloff R.G (ed), San Diego, California : College-Hill Press.

[46] Howard S. (1995), "*Intrasigent articulation disorder: using electropalatography to access and remediate misarticulated fricatives.*", In Perkins M, and Howard S. (Eds), Case Studies in Clinical Linguistics, Singular Publishing, 1995.

[47] Huckvale M., (1996), *"Speech filing system (SFS)"*, Manual, Release 3, University College of London.

[48] IBM, International Business Machines Corporation, (1992), *"IBM Speech Viewer II"*, User's Guide, page 160.

[49] Iivarinen J., Kohonen T., Kangas J., and Kaski S. (1994), *"Visualising the clusters on the self-organizing map"*, Multiple Paradigms for Artificial Intelligence (SteP94), 122-126. Finnish Artificial Intelligence Society.

[50] Katz, R.C. (1986), *"Aphasia treatment and microcomputers"*, London : Taylor & Francis.

[51] Keller E., Ostry D.J. (1983), *"Computerised measurement of tongue dorsum movements with pulsed-echo ultrasounds."*, Journal of the Acoustical Society of America ; 73:1309-1315

[52] Kershaw D.J. (1997), *"Phonetic context-dependency in a hybrid ANN/HMM speech recognition system"*, pp.33, Unpublished PhD Thesis, University of Cambridge.

[53] Kewley - Port D., Watson C.S., and Cromer P.A. (1987), *"The Indiana Speech Training Aid ISTRA: A microcomputer-based aid using speaker-dependent speech recognition"*. Synergy '87, The 1987 ASHF Computer Conference, Proceedings, pp. 94 - 99.

[54] Kewley-Port D., and Watson C.S. (1995), *"Computer Assisted Speech Training: Practical Considerations"*. In A. Syrdal, R. Bennet & S. Greenspan (Eds.). Applied Speech Technology (pp. 565-582). Boca Raton: CRC Press

[55] Kewley-Port D., Watson C.S., Elbert M., Maki K., Reed D. (1991), *"The Indiana Speech Training Aid ISTRA II : training curriculum and selected case studies"*, Clinical Linguistics and Phonetics, vol. 5.

[56] Kohonen T. (1988), *"The Neural Phonetic Typewriter"*, IEEE Computer : 11-22, March 1988

[57] Kohonen T., Hynninen J., Kangas J., Laaksonen J., and Torkkola K. (1996), *"LVQ_PAK : The learning vector quantisation program package"*, Technical report A30, Helsinki University, Laboratory of Computer and Information Science.

[58] Ladefoged P. (1982), *"A Course in Phonetics"*, Harcourt Brace Jovanovich, Publishers, London, U.K, pp 199, 208

[59] Leinonen L., Mujunen R., Kangas J., and Torkkola K. (1993), *"Self-organised acoustic feature map in detection of fricative-vowel coarticulation"*. J.Acoust.Soc.Am.,Vol. 93, No. 6, June 1993, pp 3468 - 3474

[60] Liberman A.C., and Mattingly I.G. (1985), *"The motor theory of speech perception"*, Cognition, 21, 1-36

[61] Liberman A.C., Cooper F.S., Shankweiler D.P., and Studdert-Kennedy M. (1967), *"Perception of the speech code"*, Psychological Review, 74, 431-461

[62] Lippmann R.P. (1982), *"A review of research on speech training aids for the deaf"*, in Speech and Language: advances in basic research and practise Vol 7, ed. N.J. Lass, pp. 105-133

[63] Mahshie J. (1995), *"The use of sensory aids for teaching speech to children who are deaf."*, In Spens, K.-E., and Plant, G. (Eds.) Speech Communication and Profound Deafness London: Whurr Publishers, Ltd.

[64] Maki J. (1983), *"Applications of the speech spectrographic display in developing articulatory skill in hearing impaired adults"*, in Speech of the hearing impaired: research, training and personnel preparation, (Ed.) M.N. Osberger, University Park Press, Baltimore, MD.

[65] Mao J., and Jain A. (1995), *"Artificial Neural Networks for Feature Extraction and Multivariate Data Projection"*, IEEE transactions on neural networks, VOL. 6, No. 2, pp. 296-317

[66] Masters T. (1993), *"Practical Neural Network Recipes in C++"*, Academic Press, Inc., London, U.K

[67] McDonald, E. (1964), *"Articulation testing and treatment: A sensory-motor approach"*, Pittsburgh: Stanwix House.

[68] Mermelstein P., Davis S.B. (1980), *"Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences."*, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-28, No. 4

[69] Moll KL. (1965), *"Photographic and radiographic procedures in speech research"*, ASHA Reports ; 1: 129-139.

[70] Nagayama I., Akamatsu N., and Yoshino T. (1994), *"Phonetic Visualisation for Speech Training System by Using Neural Network"*. Proc. 1994 International Conference on Spoken Language Processing (ICSLP94) pp 2027 - 2030

[71] Naher S., Uhrig C. (1996), *"The LEDA User Manual V.R.3.3"*, Manual, Max-Planck-Institut fur Informatik 66123 Saarbrucken, Germany

[72] Ohala J. (1992), *"The segment: primitive or derived ?"*. In G.J. Docherty and D.R. Ladd (Eds) Papers in Laboratory Phonology II: Gesture, Segment, Prosody. Cambridge: CUP.

[73] Öster A-M (1996). *"Clinical applications of computer-based speech training for children with hearing impairment"*. In: Proc of ICSLP-96, 4th Intl Conference on Spoken Language Processing, Philadelphia, USA, Oct 1996; 157-160.

[74] Perkell J.S, et. al. (1992), *"Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements"*, J.Acoust.Soc.Am., vol. 92, no. 6, pp. 3078-3096, 1992

[75] Perkins W.H, (1987), *"Speech pathology: an applied behavioral science"*, 2nd edition, C.V. Mosby, St Louis, MI

[76] Petrounia E.B, (1984), "*Νεοελληνική γραμματική και συγκριτική αντιπαραθετική ανάλυση"*, vol 1, Γενικές γλωσσικές αρχές - φωνητική - εισαγωγή στη φωνολογία, Part I : Theory, , pp. 295-297, Thessalonika.

[77] Popple J., and Wellington W. (1996), *"Collaborative working within a psycholinguistic framework"*, Child Language Teaching and Therapy, 12(1): 60-70.

[78] Potter R.K., Kopp G.A., and Kopp H.G. (1966), *"Visible Speech"*, New York: Dover.

[79] Povel D., Arends N. (1991), *"The Visual Speech Apparatus: Theoretical and practical aspects"*, Journal of Speech Communication, 10, pp 59-80.

[80] Povel D.J, Wansink M. (1986), *"A computer-controlled vowel corrector for the hearing impaired"*, Journal of Speech and Hearing Research, vol. 29, pp. 99-105

[81] Pratt S., Heintzelman A.T., and Deming S. Ensrud (1993), *"The efficacy of using the IBM Speech Viewer vowel accuracy module to treat young children with* *hearing impairment"*, Journal of Speech and Hearing Research, vol. 36, pp. 1063-1074.

[82] Rabiner L.R. and Schafer R.W. (1988), *"Digital Processing of Speech Signal"*, Prentice-Hall.

[83] Rauber T.W., Barata M.M., and Steiger-Garcao A.S. (1993), *"A toolbox for analysis and visualisation of sensor data in supervision"*, International Conference on Fault Diagnosis, 5-7 April, Toulouse, France.

[84] Reynolds J., and Tarassenko L. (1993), *"Learning Pronunciation with the Visual Ear"*, Neural Computing & Applications (1993) 1: pp 169 - 175

[85] Rihkanen H., Leinonen L., Hiltunen T., and Kangas J. (1994), *"Spectral Pattern Recognition of Improved Voice Quality"*. Journal of Voice : Vol.8, No. 4, pp. 320 - 326. 1994 Raven Press, Ltd., New York.

[86] Risberg A. (1968), *"The development of speech-processing aids for the deaf – past, present and future",* ESCA Workshop Speech and Language Technology for the Disabled, Stockholm, pp. 9-14

[87] Rodellar V., Nieto V., Gomez P., Martinez D., and Perez M. (1994), *"A Neural Network for Phonetically Decoding the Speech Trace"*. Proc. 1994 International Conference on Spoken Language Processing (ICSLP94) pp 1575 – 1578

[88] Rooney E., Carraro F., Dempsey W., Robertson K., Vaughan R., Jack M., Murray J. (1994), *"HARP – An autonomous speech rehabilitation system for hearing-impaired people",* Proc. 1994 International Conference on Spoken Language Processing (ICSLP94) pp 2019 – 2022

[89] Rooney E., Jack M., Lefevre J., and Sutherland A. (1995), *"HARP - A speech training aid for the hearing impaired"*, 2nd TIDE Congress, La Villette, Paris, 26th-28th April 1995

[90] Rossiter D., Howard D.M, and Downes M. (1993), *"A realtime LPC based vocal tract area display for voice development"*, Voice Foundation's 22nd Anniversary Symposium: Care of the Professional Voice, June, Philadelphia, Pennsylvania.

[91] Roy D., and Pentland A. (1998), *"A Phoneme Probability Display for Individuals with Hearing Disabilities"*, third international ACM SIGCAPH conference on assistive Technologies, Assets'98, April 15-17, CA, USA.

[92] Ruscello D. (1995), *"Visual feedback in treatment of residual phonological disorders"*, Journal of Communication Disorders, Vol. 28, No. 1

[93] Sammon J.W. (1969), *"A nonlinear mapping for data structure analysis"*, IEEE Trans. On Computer, C-18, 5, pp.401-409

[94] Sendlmeier W.F. (1995), *"Feature, Phoneme, Syllable or Word: How is speech mentally represented?"*, Phonetica, 52, 131-143.

[95] Spencer A. (1996), *"Phonology theory and description"*, Blackwell, Oxford, U.K

[96] Stackhouse J., and Wells B. (1997), *"Children's Speech and Literacy Disorders: A psycholinguistic framework"*, London: Whurr.

[97] Tuller B., Shao S., and Kelso J.A.S. (1990), *"An evaluation of an alternating magnetic field device for monitoring tongue movements"*, J.Acoust.Soc.Am., vol. 88, no. 2, pp. 674-679

[98] Van Riper, C. (1939), *"Speech correction: Principles and methods",* (1st ed) Englewood Cliffs, NJ: Prentice-Hall.

[99] Watson C.S., Kewley-Port D. (1989), *"Advances in Computer-based speech training (CBST): Aids for the profoundly hearing impaired"*, in Research on the Use of Sensory Aids for Hearing-Impaired Persons, N. McGarr, (Ed.), Volta Review, 91(4), 29-45.

[100] Wrench A.A., Jackson M.S., Jack M.A., Soutar D.S., Robertson A.G., MacKenzie J., and Laver J. (1993), *"Speech Therapy Workstation for the Assessment of Segmental Quality: Voiceless Fricatives"*, Proc. EUROSPEECH 93, Berlin, Vol. 1, pp 219-222

[101] Youdelman K., MacEachron M., and Behrman A. (1988), *"Visual and tactile sensory aids: Integration into an ongoing speech training program"*, Volta Review, vol. 1, p. 198.

[102] Young S.J., Woodland P.C., Byrne W.J. (1993), *"HTK: Hidden Markov Model Toolkit V1.5"*, Cambridge University Enginnering Department, Speech Group and Entropic Research Laboratories Inc.

[103] Young S.J., Woodland P.C., Odell J., Ollason D. (1995), *"The HTK Book, Draft"*, Cambridge University Engineering Department, Speech Group and Entropic Research Laboratories Inc.

[104] Zahorian S.A, Jagharghi A.J. (1992), *"Minimum mean square error transformations categorical data to target positions"*, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 40, pp. 13-23

[105] Zahorian S.A, Venkat S. (1990), *"Vowel articulation training aid for the deaf"*, Proc. ICASSP '90, pp. 1121-1124

[106] Zell A., Mamier G., Vogt M., Mache N., Hubner R., Doring S., Herrmann K., Soyez T., Schmalzl M., Sommer T., Hatzigeorgiou A., Posselt D., Schreiner T., Kett B., and Clemente G. (1996), *"Stuttgard Neural Network Simulator (SNNS)"*, User Manual, Version 4.0, Institute for parallel and distributed high performance systems (IPVR), University of Stuttgart,

[107] Ziegler W., Vogel M., Teiwes J., and Ahrndt T. (1997), *"Microcomputer-Based Experimentation, Assessment and Treatment"*, In Ball M.J., and Code C. (Eds), <u>Instrumental Clinical Phonetics</u>, Whurr Publishers

[108] Zimmer A.M., Dai B., and Zahorian S.A. (1998), *"Personal computer software vowel training aid for the hearing impaired"*, International Conference on Acoustics, Speech and Signal Processing, May 12-15, Seattle, Washington, USA.

[109] Zou M. (1999), *"The EZ Widget and Graphics Library V.1.38"*, Manual, Department of Mathematics, The University of Texas at Austin, TX 78712

Appendices

I Definitions

<u>(Articulation)</u> - “Articulation is the process by which sounds, syllables, and words are formed when your tongue, jaw, teeth, lips and palate alter the air stream coming from the vocal folds.”

Source : American Speech-Language Hearing Association (ASHA)
An ASHA brochure Questions and Answers about Articulation Problems
http://www.kidsource.com/ASHA/articulation.html

<u>(Articulation)</u> - “Articulation is the process through which the individual speech sounds are produced by the tongue, jaw, lips, and soft palate acting on the outgoing breath stream or vocal tone.”

Source : (Bloodstein 1984, **[12]**)

<u>(Articulation)</u> - “The process of articulation constitutes the formation of the amplified sound into words, through movements of the lips, tongue, and soft palate of the mouth, and of the related facial muscles.”

Source : Grolier Electronic Publishing, © 1992 Inc.

<u>(Articulatory disorder)</u> - “Articulation is defective when phonemes are perceived as omitted, substituted or distorted”

Source : (Perkins 1987, **[75]**)

<u>(Articulatory disorder)</u> - “An articulation error, or disorder, is a non-standard production of one or more speech sounds. There are three basic types of articulatory defects: omissions, substitutions … and distortions”

Source : (Emerick and Hatten 1989, **[21]**)

(Articulatory disorder) "…use articulatory disorder for pathologies in which the vocal organs are impaired."
Source : (Hewlett 1985, **[43]**)

(Phonetics) – "Phonetics is the study of the sensible manifestation of language. It is therefore concerned with the acoustical properties of speech, with the motor behaviour of the vocal organs that produce the acoustical signal, and with the way the signal is processed in the human auditory system."
Source : (Henderson 1971, **[42]**)

(Phonetics) – "The study of the total range of speech sounds that can be made by human beings is the concern of GENERAL PHONETICS. All human beings have substantially the same speech apparatus, so that the total repertory of human sounds is effectively the same for the whole species… The study of speech as a universal human phenomenon is PHONETICS."
Source : (Halle & Stevens 1979, **[33]**)

(**Phonology)** – "The study of the systematic organisation of selected speech sounds in the spoken forms of individual languages has variously been called FUNCTIONAL PHONETICS, PHONEMICS or more commonly nowadays, PHONOLOGY.
Source : (Henderson 1971, **[42]**)

(Phonological disorder) "…a disorder involving the phonological representations of words in the speaker's brain, or the mental processes used in the conversion of phonological forms into phonetic forms."
Source : (Hewlett 1985, **[43]**)

(Phonological – Phonetic – Articulatory level) "…phonetic level requires that an utterance consisting of a series of speech sounds be realised as a sequence of constantly changing combinations or articulatory gestures in time. So the speaker has to learn not only how to implement the gestures associated with individual speech sounds but also how they may affect or be affected by other articulatory gestures with which they may be combined. It is possible that these learned skills of phonetic implementation may become impaired independently of, or together with, the learned information concerning the phonological categories themselves, on the one hand, and some impairment to the neurological or organic structure of the vocal tract, on the other."

Source : (Hewlett 1985, **[43]**)

(Phonetic disorder) "… a three way distinction is required, including a term that describes disorders affecting a stage of linguistics processing intermediate between the phonological and the articulatory, a stage which is concerned particularly with the control and relative timing of articulatory movements."

Source : (Hewlett 1985, **[43]**)

II A Draft Manual of OLTK Interface

A. OLTK main window

The application starts with one window, which we call the 'main window' (**Figure Appendices-1**). This is split into three parts, the menu bar, the graphics canvas, and the status bar.

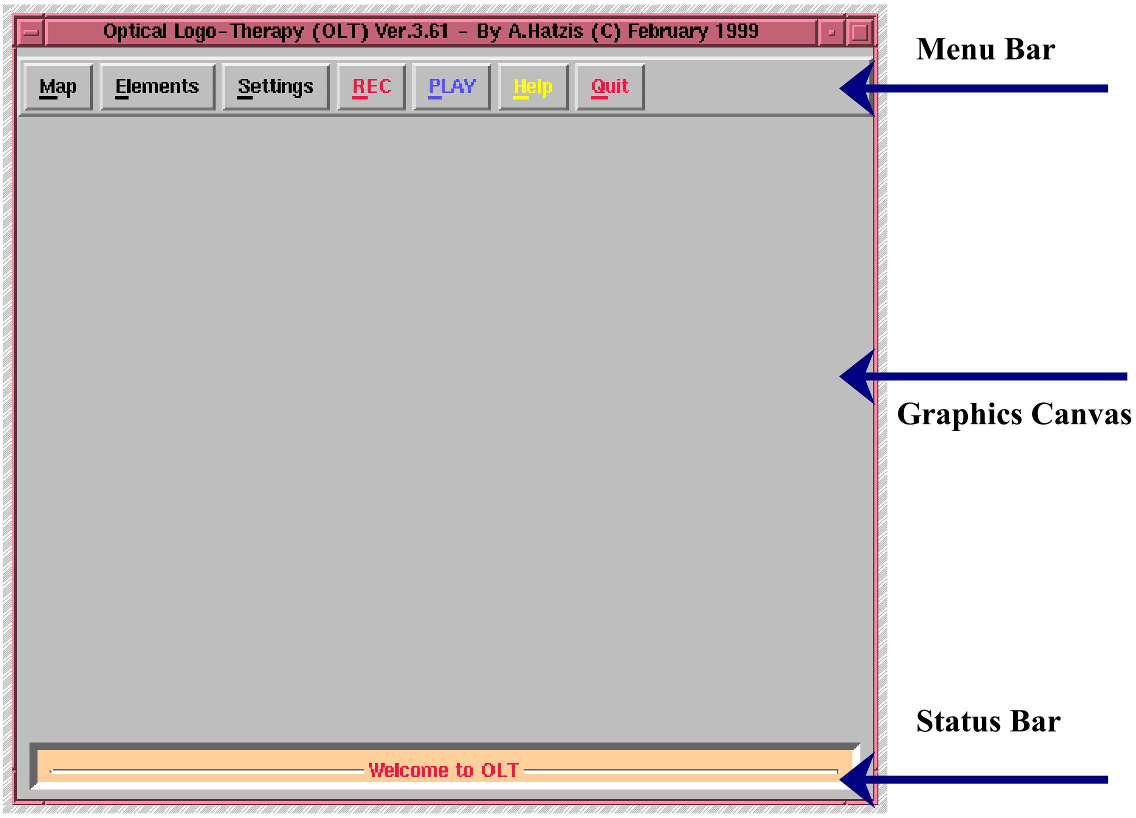


Figure Appendices-1 *OLTK main window*

a) The menu bar

All options of OLTK application are accessible through keystrokes or mouse-button clicks. The menu options have been designed in the same style that all graphical user interfaces follow nowadays. In practice this means that there are the following button types (**Figure Appendices-2**) :

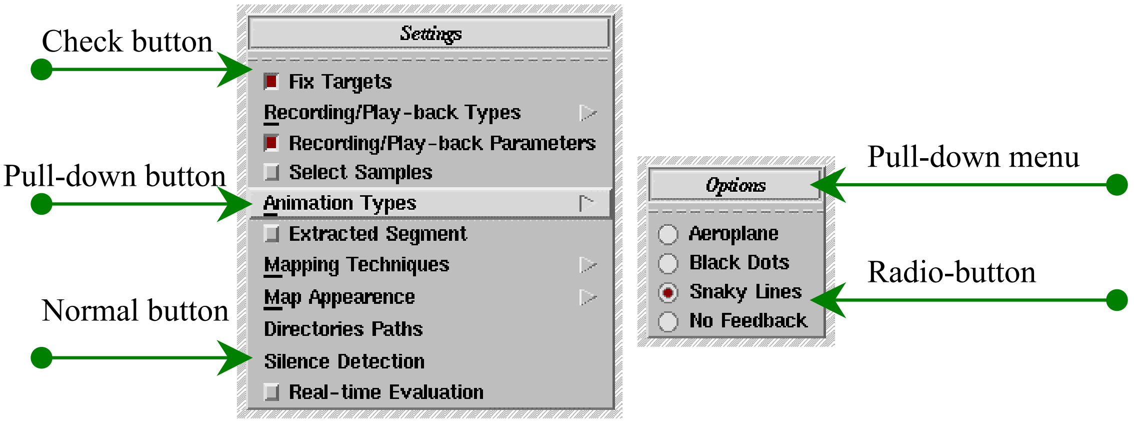


Figure Appendices-2 *Different types of OLTK buttons*

Normal button : **Executes a certain program routine**

Pull-down button : **Serves as a hook to other menu buttons.**

Check button : **Toggles between two choices.**

Radio button : **Selects from a group of mutually exclusive choices.**

The many options of the program are categorised under six menu buttons that appear on the top of the main window (**Figure Appendices-1**). All of them are 'pull-down' buttons except the 'REC' and 'PLAY' which are 'normal' and are related to recording and play-back processes respectively.

b) The graphics canvas

The largest part of OLTK main window is occupied by the graphics canvas, located under the menu bar (**Figure Appendices-1**). This is the area where the main activities of the software application take place. These activities include :

- **The different types of visual feedback in the form of :**
  - *Animation of a sprite – the 'aeroplane' game*-**Figure Appendices-9**.
  - *Drawing points on the map – the 'black dots' game* ***-Figure 5-3.***
  - *Drawing trajectories – the 'snaky lines' game* ***-Figure 5-3***.
- **The drawing of phonetic maps in the form of phones or clusters.**
- **The 'acceptance', 'rejection', 'silence' indicator on the face of a clown.**
- **The reward given to the user for successful efforts in the form of stars.**

- **The design of a new phonetic map by placing targets in appropriate 2D positions.**

c) <u>The status bar</u>

On the bottom part of the window (**Figure Appendices-1**) we can notice the status bar. The status bar was designed to provide the user with diagnostic messages concerning errors happening during the run-time of the program. These messages about errors are displayed immediately in the terminal window. The status bar is also used to display 'help' messages on several of the functions of the software application and on the access to menu options. It is possible for example to see the mouse co-ordinates and how these are related to the world co-ordinates of a point inside the drawing area of the graphics canvas.

B. Map operations

a) <u>Map options</u>

These are grouped under the 'Map' main menu button and include the options 'Load Map', 'Show Map', 'Save Map' and 'Create Map' (**Figure Appendices-3**). The most important of these in the current version of OLTK is the 'Load Map' option.

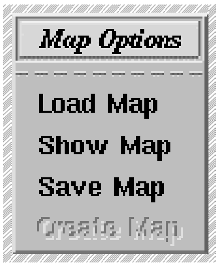
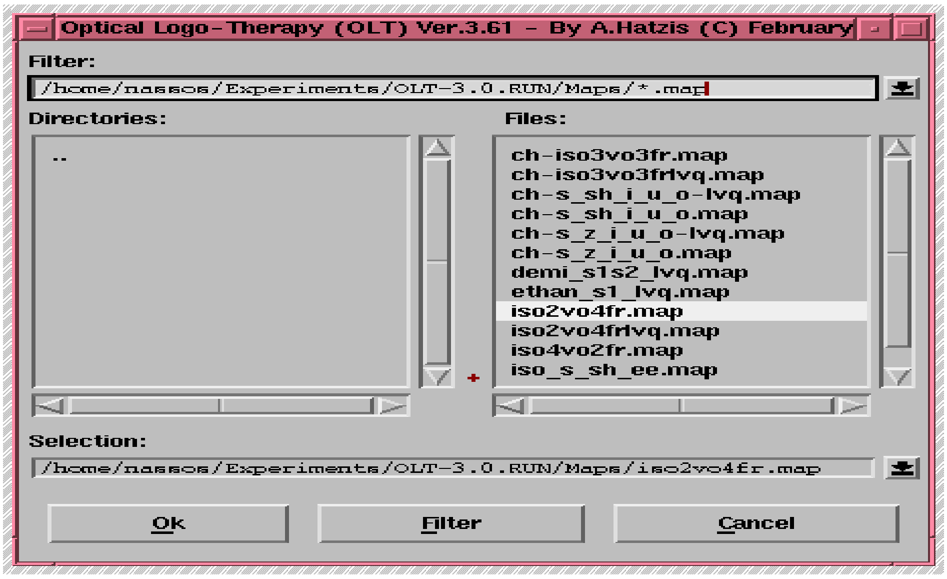


Figure Appendices-3 *Main menu – Map options*

Load Map

Once OLTK has started, the first action of the user is to load an existing phonetic map. This option also overwrites the previous loaded map and returns to the initial default settings. The only user action involved in loading a map is to select the file through the 'File Selector' window (**Figure Appendices-4**). This is a filtered list of files with the '.map' extension inside a specific directory from which the user can select the map file. The technical details and other information are shown during loading time on the terminal window.


Optical Logo-Therapy (OLT) Ver.3.61 - By A.Hatzis (C) February
OK
Filter
Cancel

Filter: /home/nassos/Experiments/OLT-3.0.RUN/Maps/*.map

Directories: ..

Files:
ch-iso3vo3fr.map
ch-iso3vo3frlvq.map
ch-s_sh_i_u_o-lvq.map
ch-s_sh_i_u_o.map
ch-s_z_i_u_o-lvq.map
ch-s_z_i_u_o.map
demi_s1s2_lvq.map
ethan_s1_lvq.map
iso2vo4fr.map
iso2vo4frlvq.map
iso4vo2fr.map
iso_s_sh_ee.map

Selection: /home/nassos/Experiments/OLT-3.0.RUN/Maps/iso2vo4fr.map

Figure Appendices-4 *The 'File Selector' window of OLTK. We can see the filtered option and the full directory path of the highlighted selected file.*

Save Map

This option is useful for saving modified or newly created maps. All statistical information and sets of multi-dimensional labelled vectors for each phonetic class are saved. The file name and directory location can be chosen from the 'File Selector' window. The fixed target positions defined with the 'Fix Targets' option can also be saved. The 'Save Map' option is currently under development, and is combined with other options of OLTK that are more technically oriented.

Create Map

Currently this option is unavailable because automating the process of creating a map was out of the scope of our research. It is present just to remind us that it is essential for a fully commercial development of OLTK to include that feature. The only function which it was necessary for the instructor to use for non-automatic creation of maps was the 'Fix Targets' option.

Show Map

The 'Show Map' option redisplays the map according to the current 'Map Appearance' settings. That means that a map is represented with 'Circles', 'Phones' or 'Ellipses' graphics, explained in the following section.

b) Settings

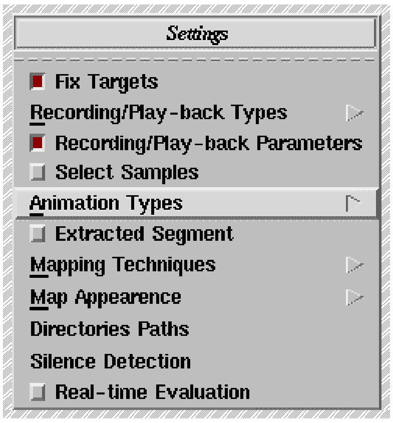
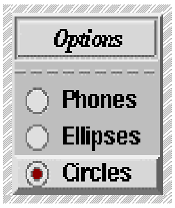


Figure Appendices-5 *Main menu – Settings options*

Map appearance



The 'Map Appearance' settings referred to in 'Show Map' options is one among the many other options of 'Settings' main menu button. The other one relevant to the map operations is the 'Fix Targets'.

Figure Appendices-6 *Main menu – Settings - Map appearance options*

Circles

Once the loading process has finished, the map is drawn according to its default appearance setting which is the 'Circles'. Each circle represents a sound cluster with its centre at the mean of the 2D distribution and the radius proportional to the 'Sensitivity' setting (**Figure Appendices-11**).

Phones

The 'Phones' option draws the map as a number of coloured, labelled boxes, each one having a specific 2D position according to the transformation (§4.6.3-II). The boxes can either represent feature vectors of 10msec speech frames or codebook vectors. In both cases we call these vectors 'phones'. All the *phones* used to model a certain sound cluster share the same colour.

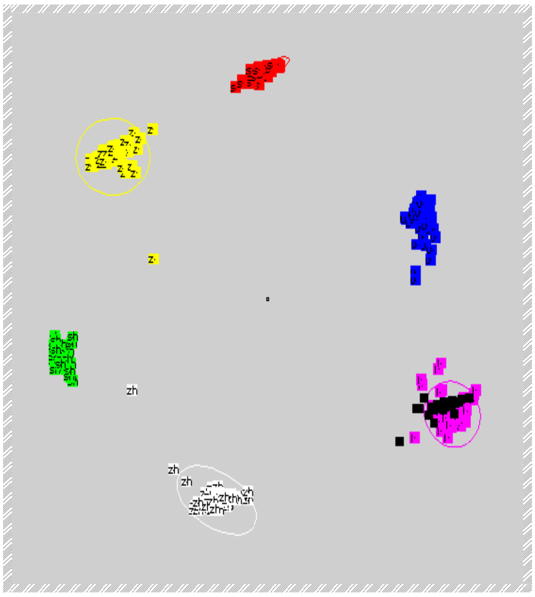


Figure Appendices-7 *The 'Phones' map representation – blackening areas.*

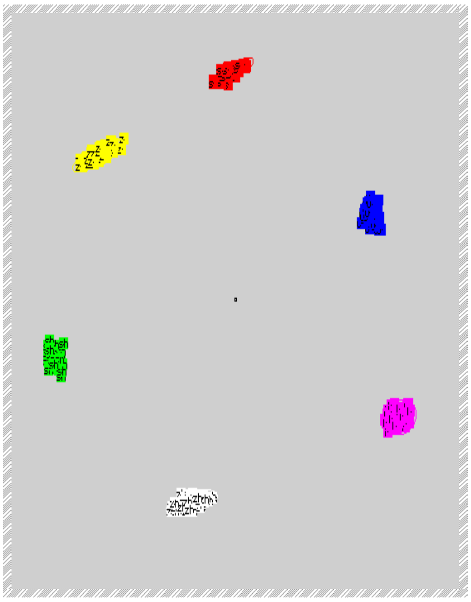


Figure Appendices-8 *Modified map after deleting certain phones.*

Ellipses

For the 'Ellipses' option, the 2D distribution of *phones* for each cluster can be modelled with a bivariate normal density, assuming Gaussian distributions. The ellipses drawn for each cluster represent an isocurve, and each one of these curves includes a percentage of the distribution. Each ellipse with its colour, variance and mean can represent a phonetic class (**Figure Appendices-9**). The centre of each ellipse coincides with the centroid of the phone cluster, and its

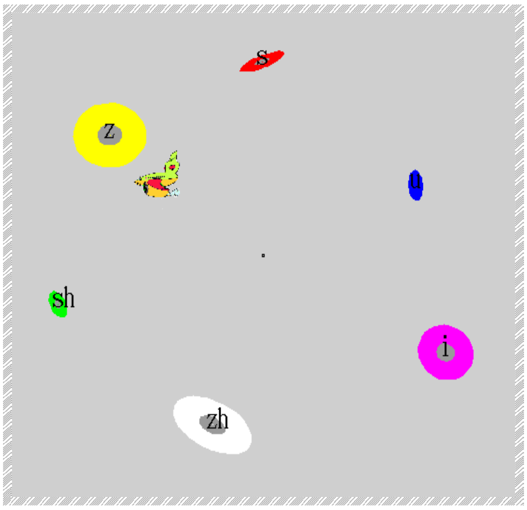


Figure Appendices-9 *The "Ellipses" map representation and the "Aeroplane" flying.*

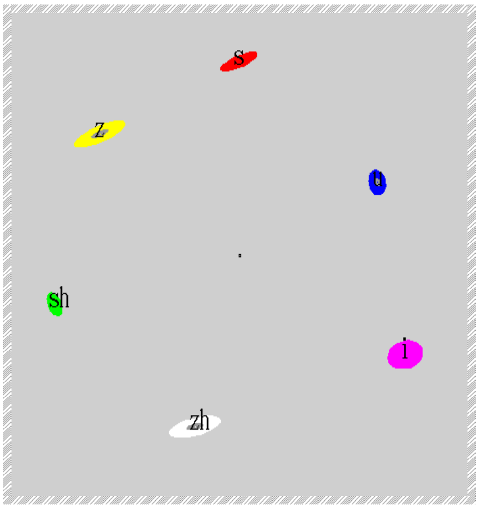


Figure Appendices-10 *Same map modified by "Edit Phones" option.*

major and minor axes coincide respectively with the direction of maximum and minimum data variation. The length of the major and minor axis is equal to three standard deviations.

Fix Targets

Activating this option from the 'Settings' menu button, displays on the graphics canvas a number of 'star' labels displaying the sound they represent (**Figure Appendices-11**). Each label is initially placed on the mean of the distribution for each sound cluster drawn on the map. The user can grab any label with the mouse and freely move it to another position. There are two main reasons this option can be helpful. First, the instructor can design the layout of the sounds to be present during speech training. Recall that the *SMANN* network learns to associate all the vectors of a specific sound class with a fixed two-dimensional position (§**4.6.3-II**). Second, it is possible to place a label near the sound cluster area to mark the position of a new two-dimensional target for the client to reach.

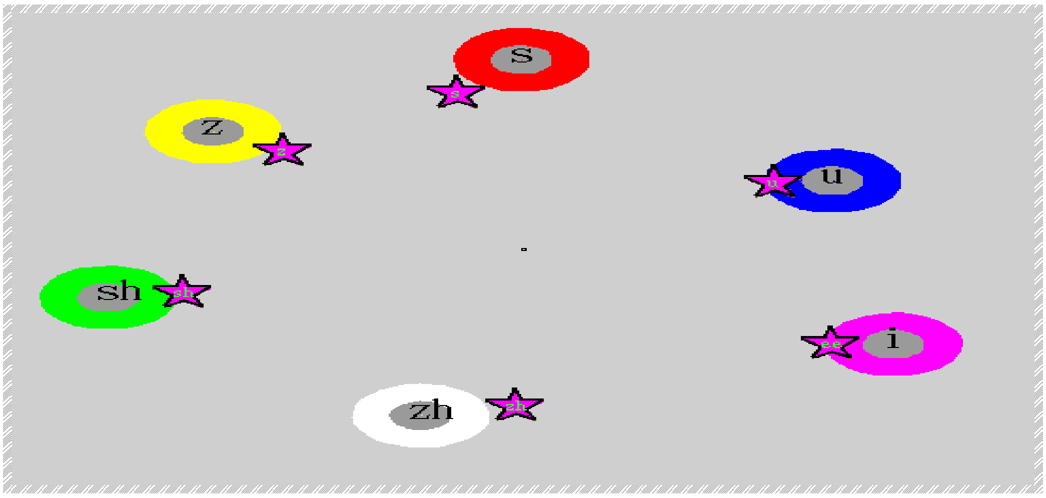


Figure Appendices-11 *The 'Circles' map representation and the 'star' labels as they are placed close to the sound clusters. The instructor can freely move them and design a map layout or provide alternative targets to the client.*

c) <u>Elements</u>

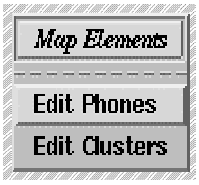
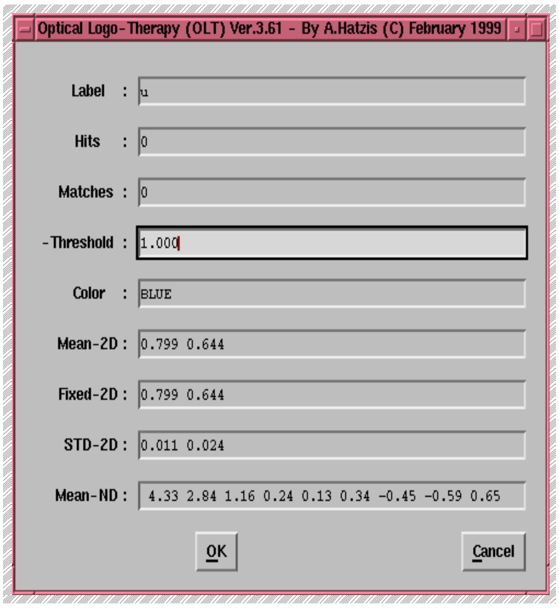


Figure Appendices-12 *Main menu – Elements options*

According to the previous discussion on the *Map Appearance*, we can differentiate between two ways of grouping and viewing the map, as *phones* and as clusters. We call these, ‘Map Elements’. The OLTK interface, provides two menu options for altering or displaying the information for each one of the *Map Elements* and consecutively modifying the phonetic map, ‘Edit Phones’ and ‘Edit Clusters’ (**Figure Appendices-12)**. These two options define also two distinct programing states, each one able to handle different *call-back* events (**§4.8.2-I**).

Edit Clusters

Once we click on the *Edit Clusters* button the map is redrawn with ellipses or with circles. Then we can click with the left mouse button on the designated coloured areas of any of the clusters and obtain information from a parameter window (**Figure Appendices-13**) : The ‘Label’ and ‘Colour’ specify the label and colour of the cluster. The ‘Hits’ specify how many frames of speech have been classified to the cluster; and if the frames of speech are labelled, the ‘Matches’ show how many of the hits have the same label as that of the cluster. The ‘Mean-2D’ and ‘STD-2D’ fields shows the mean and standard deviation of the 2D distribution of the cluster phones. The ‘Fixed-2D’ is the fixed cluster centroid



Optical Logo-Therapy (OLT) Ver.3.61 - By A.Hatzis (C) February 1999

Label : u

Hits : 0

Matches : 0

-Threshold : 1.000

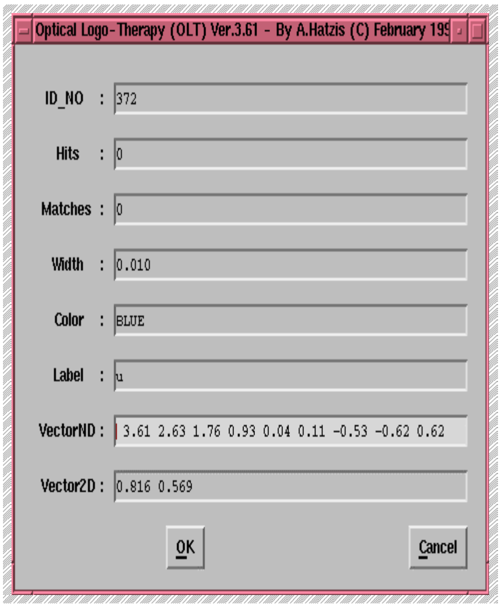
Color : BLUE

Mean-2D : 0.799 0.644

Fixed-2D : 0.799 0.644

STD-2D : 0.011 0.024

Mean-ND : 4.33 2.84 1.16 0.24 0.13 0.34 -0.45 -0.59 0.65



Optical Logo-Therapy (OLT) Ver.3.61 - By A.Hatzis (C) February 199

ID_NO : 372

Hits : 0

Matches : 0

Width : 0.010

Color : BLUE

Label : u

VectorND : 3.61 2.63 1.76 0.93 0.04 0.11 -0.53 -0.62 0.62

Vector2D : 0.816 0.569

Figure Appendices-13 *The parameters window for editing a cluster.*

Figure Appendices-14 *The parameters window for editing a phone.*

we select at the time we design the layout of the map. The ‘Mean-ND’ field has the value of the mean of the feature vectors that comprise the cluster. Finally the ‘-Threshold’ field displays the decrement threshold value of the cluster relative to the global sensitivity (**§5.3.2-D**). This is the only value we allow the user to modify in the current version of the program.

Edit Phones

Similarly, selecting the *Edit Phones*, we can click with the left mouse button on any of the coloured boxes that represent the phones of the map. A different window pops up and displays information about the phone (**Figure Appendices-14**); an identification number, ‘ID_NO’, which is the position of the phone vector inside the file, the ‘Colour’, and the ‘Label’ of the phone, the ‘Hits’ and ‘Matches’ according to how many frames of speech are close to the phone vector and how many of them have the same label as that of the phone, the ‘Width’ representing the size of the drawn coloured box and usually proportional to the *Hits* or *Matches*, and finally the ‘VectorND’ that shows the parameters of the vector represented by the box drawn at the position specified by the ‘Vector2D’ field. In the present version of OLTK parameters cannot be modified. Despite those limitations, user can click with the right mouse button on any phone and delete it from the map. The ellipses are recalculated and the map is redrawn (**Figure Appendices-10**) this way the user can modify the 2D phone cluster distribution and save the map with the *Save Map* option.

C. Recording/Play-back operations

a) <u>REC/PLAY menu buttons</u>

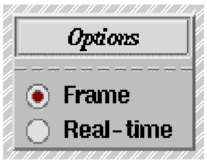
On pressing either of these buttons the indicated mode of operation starts immediately. In the *REC* mode we can only record a certain utterance in real-time. In the *PLAY* mode, however, there are many variant forms for play-back routine. So we have play-back of a previously recorded utterance, or of a segment extracted from an utterance, or of previously saved utterances. A special play-back mode for playing frames of speech is also available **(§5.3.2-B)**.

Moreover, in those modes the user can press certain keyboard or mouse buttons that correspond to other special operations (**Table Appendices-1**).

List of operations in the REC/PLAY mode of OLTK	
Mouse buttons and ESC	
Left	Redraws the graphics canvas
Right	Ends the recording or play-back
ESC	Ends the recording or play-back
Arrow keys for controlling the '*Frame Play-back Tool*'	
Down	Plays forward a sequence of speech frames
Up	Plays backward a sequence of speech frames
Left	Stops the play-back of speech frames
Right	Stops the play-back of speech frames
Function keys for other displays and functions	
F6	Save utterance
F7	LYRE time domain waveform display
F8	SFS spectrogram and time domain waveform display
F9	GNUPLOT graph of time vs. threshold distances
F10	Extract a speech segment
F11	GNUPLOT graph of % of accepted frames of speech vs. threshold
F12	GNUPLOT graph of utterance acceptance vs. time

Table Appendices-1 *Extra user options by pressing keyboard and menu buttons*

b) <u>Rec/Play Types</u>



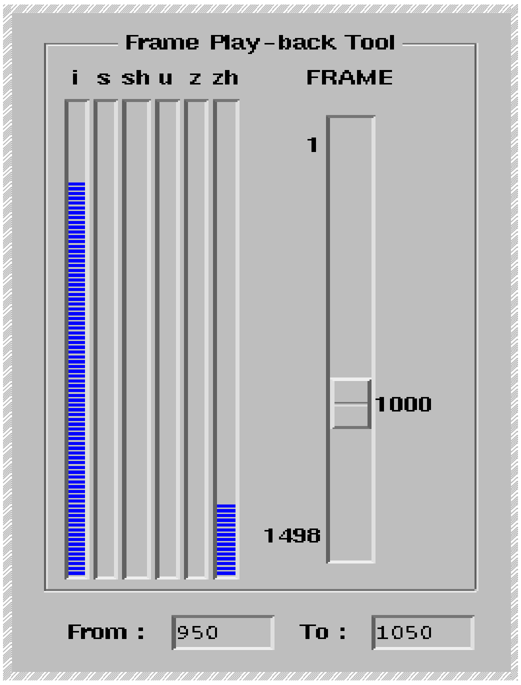
There are two radio button options in the 'Rec/Play Types' pull-down menu, 'Frame' and 'Real-Time'.

Figure Appendices-15 *Main menu – Settings - Rec/Play types options*

'Frame-by-frame' analysis

This option works only for the *PLAY* mode. There is meant to be a similar mode of *REC* for recording a predefined sequence of utterances, but currently is not available in OLTK. To exemplify the operation of *Frame* option we recorded some speech and illustrated it with three different time-aligned displays obtained by pressing F8, F9, and F12 (**Figure Appendices-19, Figure Appendices-20, Figure Appendices-21**).

When the *radio* button of the *Frame* option is activated and we press the *PLAY* button the ‘Frame Play-back Tool’ pops (**Figure Appendices-16**). On the left-hand side there is a bar-graph of the sound classes and the output drawn is the posterior classification of our neural network classifier. The height of the bar shows how accurately the incoming speech frame is classified. In our aforementioned example, a frame of sound /i/ on the 10th second of speech is classified mostly as /i/ but it is also confused a bit with /ʒ/ (**Figure Appendices-16**). This is in accordance here with the 2D mapping of the frame; note that the mapping is relevant to the classification result, where the position is closer to /i/ and further away from /ʒ/ (**Figure Appendices-17**).



From : 950 To : 1050

Figure Appendices-16 *A frame of the utterance /si/ classified*

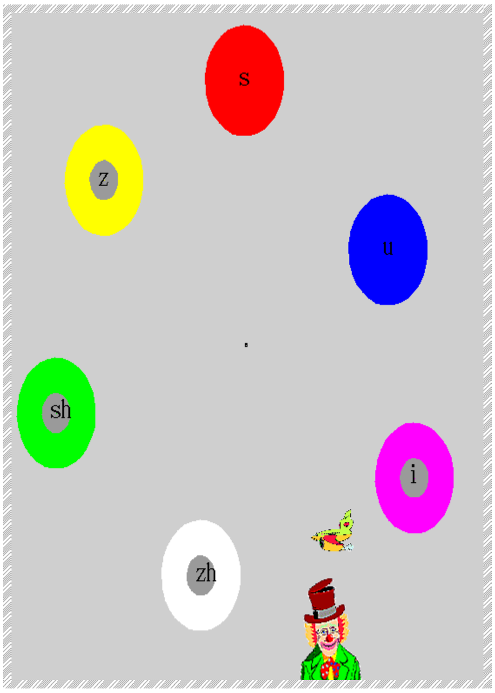
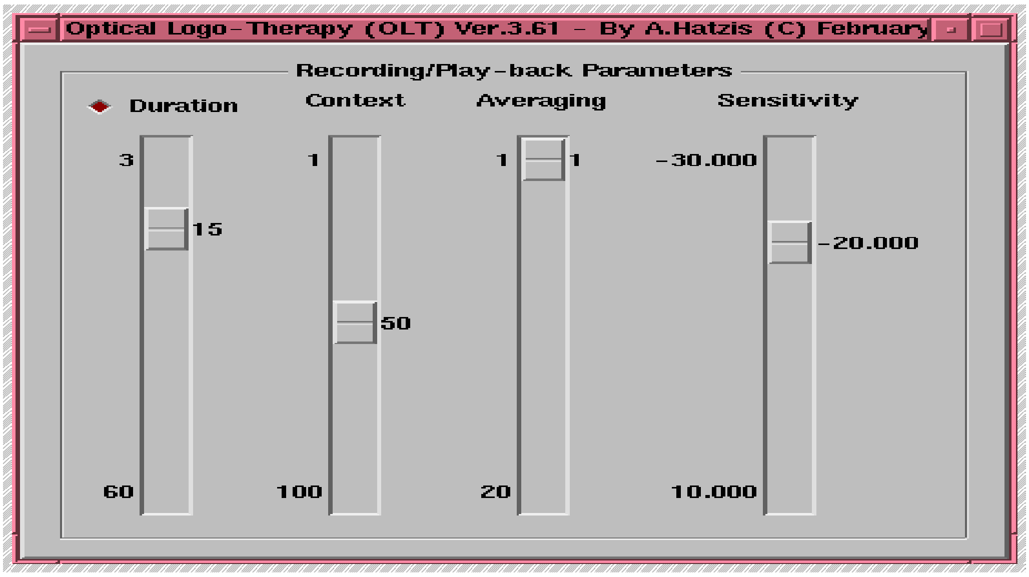


Figure Appendices-17 *The mapping of the 1000th frame of utterance /si/*

On the right-hand side of the frame play-back tool we see a slider. This slider can be positioned on the exact frame of speech we wish to examine and also it can be moved backward and forward resulting in animation of a specific speech segment or of the whole utterance on the graphics canvas. Finally, the two numbers at the bottom of the *'Frame Play-back Tool'* specify the width of a segment of the utterance we examine centred in the current position of a speech frame.

c) 'Rec/Play' Parameters



Optical Logo-Therapy (OLT) Ver.3.61 - By A.Hatzis (C) February

Recording/Play-back Parameters

◆ Duration: 15 (range 3 – 60)

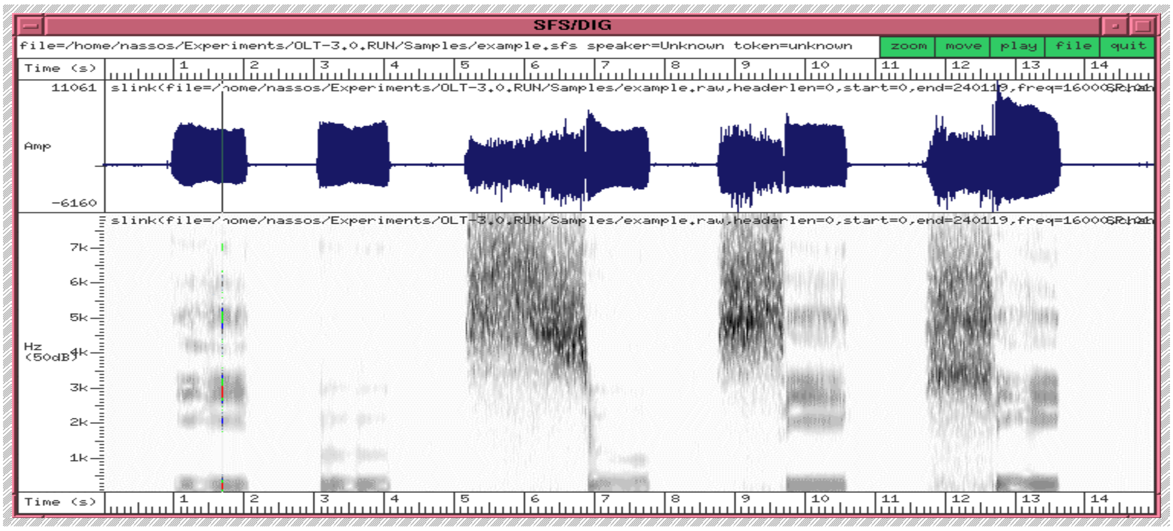
Context: 50 (range 1 – 100)

Averaging: 1 (range 1 – 20)

Sensitivity: -20.000 (range -30.000 – 10.000)

Figure Appendices-18 *Main menu – Settings - Recording/Playback Parameters – Sliders activated from the radio button option*

These are four : duration, context, averaging and sensitivity. All of them can be controlled with slide-bars (**Figure Appendices-18**). For the explanation and use of them see §**5.3.2**.



file=/home/nassos/Experiments/OLT-3.0.RUN/Samples/example.sfs speaker=Unknown token=unknown

slink(file=/home/nassos/Experiments/OLT-3.0.RUN/Samples/example.raw,headerlen=0,start=0,end=240119,freq=16000,SR=...

Figure Appendices-19 *SFS display of speech. The utterances recorded are /i/, /u/, /su/, /si/ and lateralised /si/. Notice the extra formant of the lateral /si/ around 3kHz.*

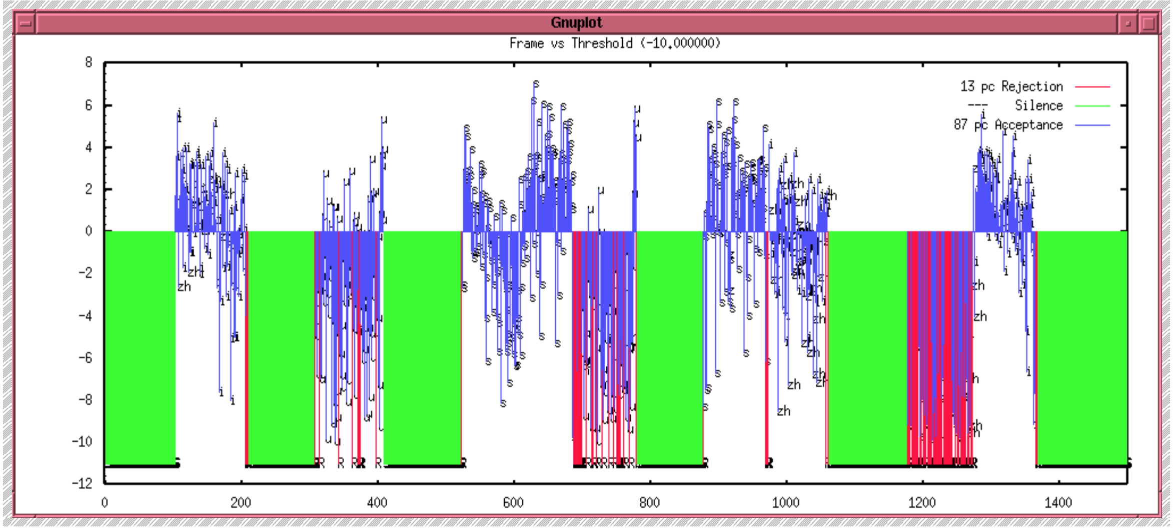


Figure Appendices-20 *Gnuplot display of time vs frame's threshold distance. Silence is drawn green, rejection is drawn red. Notice how the frames of lateral /si/ are rejected.*

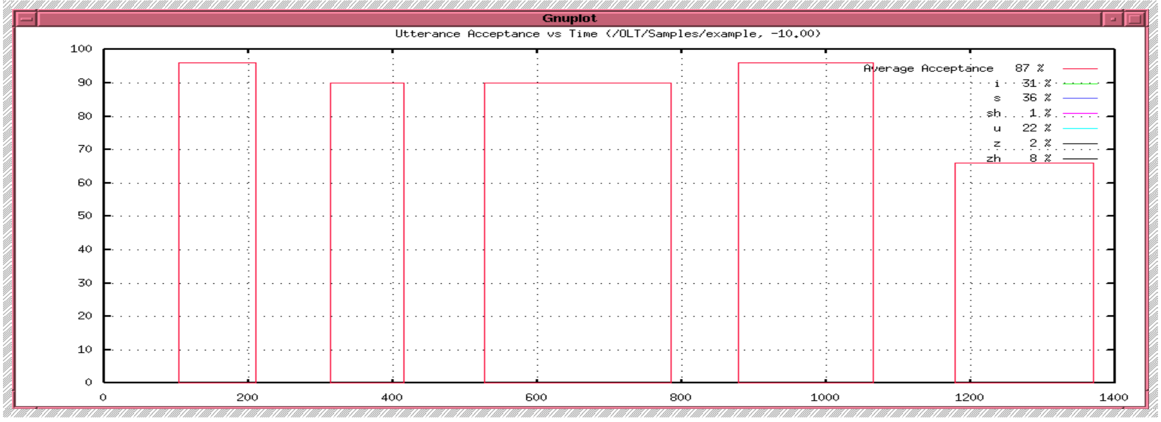
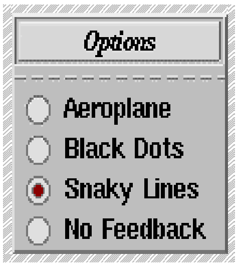


Figure Appendices-21 *Display for the % of acceptance for each utterance and for each sound class.*

D. Other operations

We discuss here the remaining options of OLTK, many of which, such as setting directories paths and silence detection parameters, are independent of the other features of the application.

a) <u>Animation types</u>



The appearance of the visual feedback produced with OLTK can be based on any of the three animation types, *Aeroplane*, *Black Dots*, and *Snaky Lines*. All of them have been designed so that they are interesting, especially for children, and serve game oriented strategies.

Figure Appendices-22 *The radio button options of 'Animation Types' pull-down menu as it appears on the 'Settings' menu button.*

Aeroplane

The game task here is to fly the aeroplane close to a cloud represented by circles or ellipses (**Figure Appendices-9**). When the sound produced passes the threshold level of the sensitivity parameter, the plane is drawn at the 2D co-ordinates of the transformed vector frame and the acceptance is indicated with the smiling face of the clown (**Figure Appendices-17**). On the contrary if the sound is rejected, the plane disappears from the map and this causes the clown to frown. Finally there is a third speech state, that of silence, which is pictured with a 'snoring' plane and a sleepy clown (**Figure Appendices-23**).

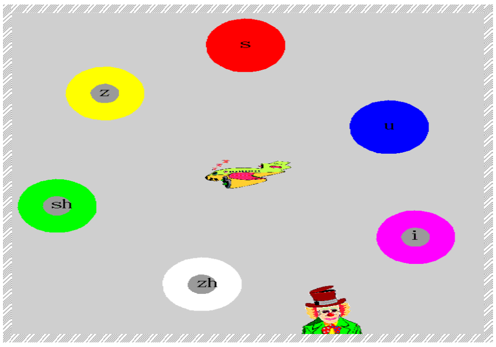


Figure Appendices-23 *The silence speech state*

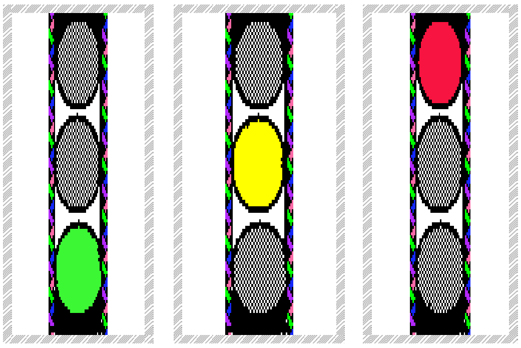


Figure Appendices-24 *The traffic lights*

Black Dots

The *Black Dots* is the game where the student's goal is to cover a certain area on the map with black dots (**Figure Appendices-7**). This type of animation and the *Snaky Lines* are better suited to visual comparison between an example articulation by the tutor and an attempted articulation by the trainee as the trace of speech is drawn on the map. The user of OLTK can redraw the map, if it is cluttered, by pressing the left mouse button. The *Black Dots* is also useful for checking the quality of mapping and whether it is consistent or not.

Snaky Lines

The third option for animation is the *Snaky Lines*. This game has been inspired by a similar video game. In our version the head of the snake is the most recent speech frame plotted on the map, and the tail of the snake is comprised at the previous ten frames of speech (**Figure 5-3-D1,D2**). The snake moves its head and tail representing speech trajectories. The visual feedback provided with the *Snaky Lines* can be most beneficial in speech drills with words or connected sounds.

No feedback

We have also included a fourth option, 'No feedback', to cover the cases where we do not wish to provide visual feedback. Such cases can be used when collecting speech samples for analysis and assessment or when we want to see how the client performs without receiving any feedback. It is important for post-treatment review to check whether students have learned the new speech skills and can self-judge and self-monitor their articulation.

Traffic lights

This is not really an animation option but it involves the animation of a traffic light in order to signal the beginning and end of a recording session (**Figure Appendices-24**). In practice it draws the attention of the speaker and prepares him/her for the task which is to be accomplished.

b) <u>Real-Time evaluation</u>

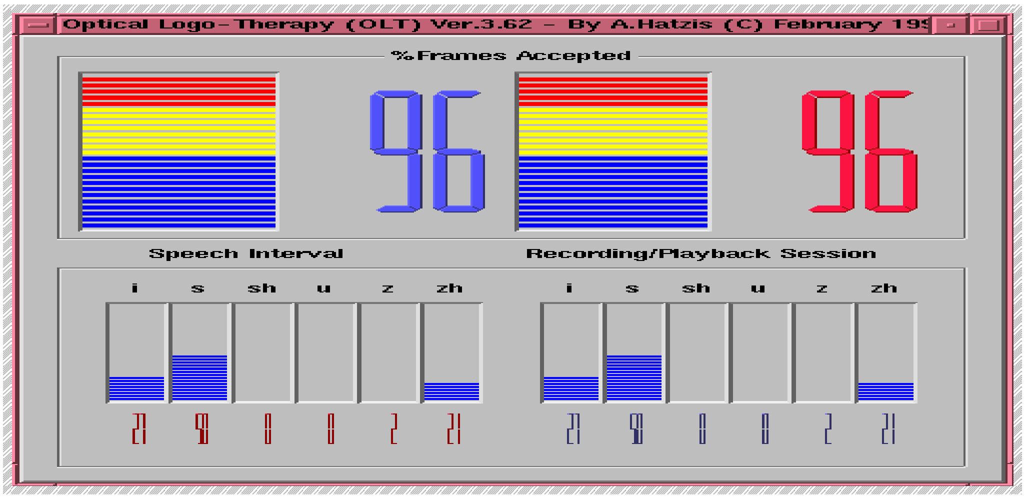

Optical Logo-Therapy (OLT) Ver.3.62 - By A.Hatzis (C) February 19
% Frames Accepted
96
96
Speech Interval
Recording/Playback Session
i
s
sh
u
z
zh
i
s
sh
u
z
zh
21
0
0
2
21
21
0
0
2
21

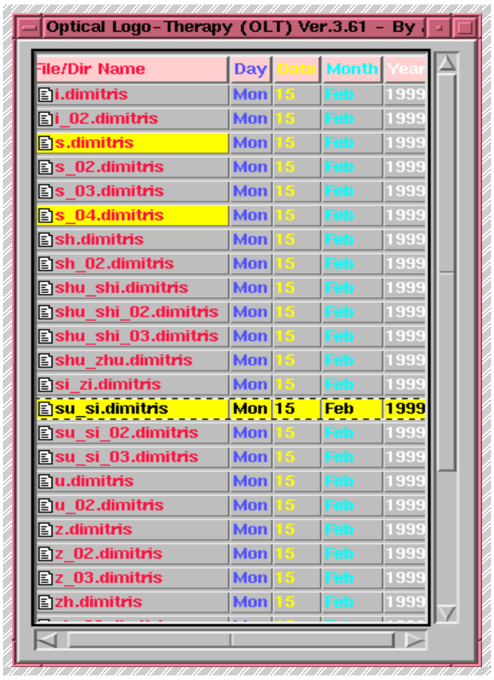
Figure Appendices-25 *Real Time evaluation display*

The display **(Figure Appendices-25)** can show results while recording or play-back speech. It is split into four parts. The upper half is for measuring the overall acceptance of frames and the lower half is for showing classification percentages for the sound accepted. The left half is for evaluation of a speech segment between two silence periods and the right half is for evaluating an average of all the spoken utterances. The scoring is shown both with big numbers and with coloured bar-graphs.

c) <u>Select Sample</u>

This check-button of the *Settings* main menu brings up another display of OLTK which contains all the speech samples stored inside a specific directory (**Figure Appendices-26**). The user can select any number of samples and use the *Play* button to play-back the sequence of files. The corresponding events are produced and the teacher or the student can see the differences or similarities of the samples loaded. In the example of **Figure Appendices-27** we can see two different articulations of the /s/ sound produced by loading two speech samples of a Greek native speaker. One of the samples is mapped closer to the /ʃ/ area while the other hit the /s/ target. In this way it is possible for the teacher to select appropriate files to demonstrate specific articulatory configurations to the student, and give instructions to match the animation and the scores of the loaded samples with his/her own attempts. Whenever the tutor wishes to save a recorded sample of the trainee for later comparison or analysis s/he can press the

F6 button to bring up a display similar to the *File Selector* (**Figure Appendices-4**). After the name of the file is given s/he can use the *Select Samples* display to compare it with other pre-recorded samples.



Optical Logo-Therapy (OLT) Ver.3.61 - By

ile/Dir Name	Day	Date	Month	Year
i.dimitris	Mon	15	Feb	1999
i_02.dimitris	Mon	15	Feb	1999
s.dimitris	Mon	15	Feb	1999
s_02.dimitris	Mon	15	Feb	1999
s_03.dimitris	Mon	15	Feb	1999
s_04.dimitris	Mon	15	Feb	1999
sh.dimitris	Mon	15	Feb	1999
sh_02.dimitris	Mon	15	Feb	1999
shu_shi.dimitris	Mon	15	Feb	1999
shu_shi_02.dimitris	Mon	15	Feb	1999
shu_shi_03.dimitris	Mon	15	Feb	1999
shu_zhu.dimitris	Mon	15	Feb	1999
si_zi.dimitris	Mon	15	Feb	1999
su_si.dimitris	Mon	15	Feb	1999
su_si_02.dimitris	Mon	15	Feb	1999
su_si_03.dimitris	Mon	15	Feb	1999
u.dimitris	Mon	15	Feb	1999
u_02.dimitris	Mon	15	Feb	1999
z.dimitris	Mon	15	Feb	1999
z_02.dimitris	Mon	15	Feb	1999
z_03.dimitris	Mon	15	Feb	1999
zh.dimitris	Mon	15	Feb	1999

Figure Appendices-26 *Selection of speech samples.*

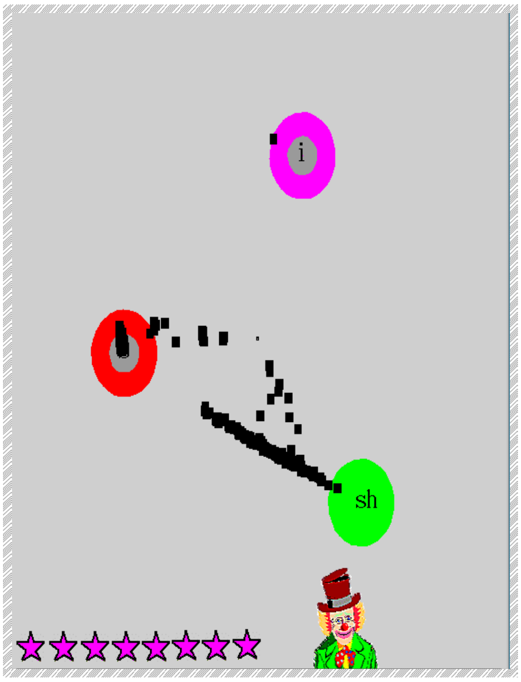
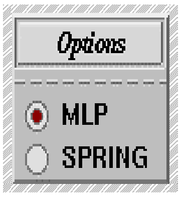


Figure Appendices-27 *Visual comparison of two loaded speech files.*

d) Mapping techniques



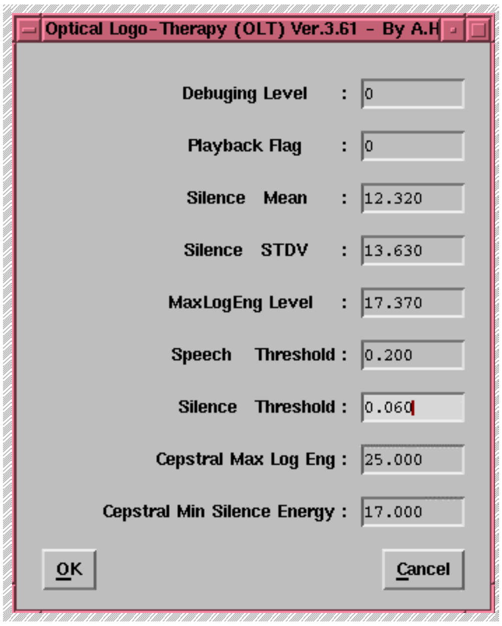
Options

● MLP

○ SPRING

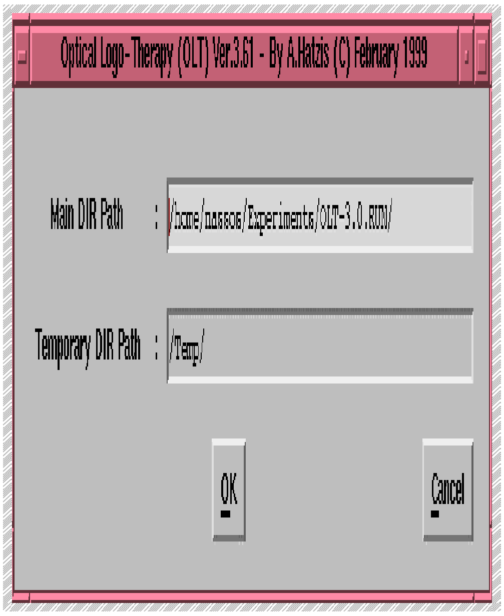
Figure Appendices-28 *The 'Mapping Techniques' options*

There are two transformations from the multi-dimensional feature vectors to points on the map, the 'MLP' and the 'Spring' (**Figure Appendices-28**). The visual feedback produced can differ significantly between the two, and it is up to the instructor and the researcher to judge which is more appropriate in what type of problems. Nevertheless according to the discussion we had in **§4.6.3-I** the mapping based on the *Spring* model is not so smooth as the *MLP* transformation.



Debuging Level : 0
Playback Flag : 0
Silence Mean : 12.320
Silence STDV : 13.630
MaxLogEng Level : 17.370
Speech Threshold : 0.200
Silence Threshold : 0.060
Cepstral Max Log Eng : 25.000
Cepstral Min Silence Energy : 17.000

Figure Appendices-29 *The display for modifying the parameters of silence detection program*


Optical Logo-Therapy (OLT) Ver.3.61 - By A.Hatzis (C) February 1999
OK
Cancel

Main DIR Path : /home/nassos/Experiments/OLT-3.0.RUN/
Temporary DIR Path : /Temp/

Figure Appendices-30 *Options to specify main and temporary directory paths*

e) Silence detection

Silence detection is critical for the correct operation of recording and play-back. The environmental conditions and the type of hardware inside the computer may change; therefore it was necessary to provide a display with all the parameters of our silence detection program, (**Figure Appendices-29**, **§4.3**). The user can modify any of these parameters and change the behaviour of the silence detection algorithm according to the conditions we referred to.

f) Directories paths

This option allow the user to specify the main directory path under which reside all the other directories for selecting, speech samples, maps, and pictures needed for OLTK (**Figure Appendices-30**). The other important directory to specify is the one for recording speech. It is important to know this so to avoid problems due to the big size of the recorded file and the limited storage capacity of the hard-disk.

g) Help - Quit

Help is currently not available; as an alternative we have created a hyper-text manual of OLTK together with some tutorial examples to practise. The *Quit* button exits the application.

III Run-Time Output of Programs

A. Program sildetect – Silence detection

```
Silence Detector machine -- usage:
sildetect [-d debug] [-p play] [-m mean] [-s stdv] [-e maxeng] [-a thresh] [-b thresh]
     [-i file] [-o file] [-h help]
where:
   -d debug       Debug Level      (default 0)
   -p play        write to /dev/dsp        (default 0 false)
   -m mean        Mean of silence          (default -1.0)
   -s stdv        STDV of silence          (default -1.0)
   -e maxeng      Maximum log energy       (default -1.0)
   -a threshold   Speech  threshold        (default 0.20)
   -b threshold   Silence threshold        (default 0.06)
   -i file        Input  filename          (default stdin)
   -o file        Output filename          (default stdout)
   -h help        Command usage
```

Example : rec | sildetect -d 1 -p 0 -o check.raw

```
Quiet Please - measuring silence
Silence measurement completed
Silence Levels Set : Offset 348.501556, Mean 12.518957, Deviation 13.763402
Say a loud utterance ....
Max Log Energy measurement completed
Max Log Energy Set : MaxLogEng = 17.635256
Silence Threshold      =   0.060      Speech  Threshold      =   0.200
Mean of silence        =  12.519      STDV of silence        =  13.763
Maximum log energy     =  17.635      PLAYFLAG               =       0
Time :    0 msec - LogEng : 16.791 - NormLogEng :  0.916 - Thresh :   0.310 - Type 2
Time :   60 msec - LogEng : 16.888 - NormLogEng :  0.925 - Thresh :   0.317 - Type 2
Time :  120 msec - LogEng : 16.852 - NormLogEng :  0.922 - Thresh :   0.315 - Type 2
Time :  180 msec - LogEng : 16.726 - NormLogEng :  0.909 - Thresh :   0.306 - Type 2
Time :  240 msec - LogEng : 15.386 - NormLogEng :  0.775 - Thresh :   0.208 - Type 2
Time :  300 msec - LogEng : 13.281 - NormLogEng :  0.565 - Thresh :   0.055 - Type 0
Time :  360 msec - LogEng : 11.728 - NormLogEng :  0.488 - Thresh :  -0.057 - Type 0
Time :  420 msec - LogEng : 12.990 - NormLogEng :  0.536 - Thresh :   0.034 - Type 0
Time :  480 msec - LogEng : 12.142 - NormLogEng :  0.488 - Thresh :  -0.027 - Type 0
```

B. Program HCodeRT – Cepstral analysis

USAGE: HcodeRT [options] infile outfile

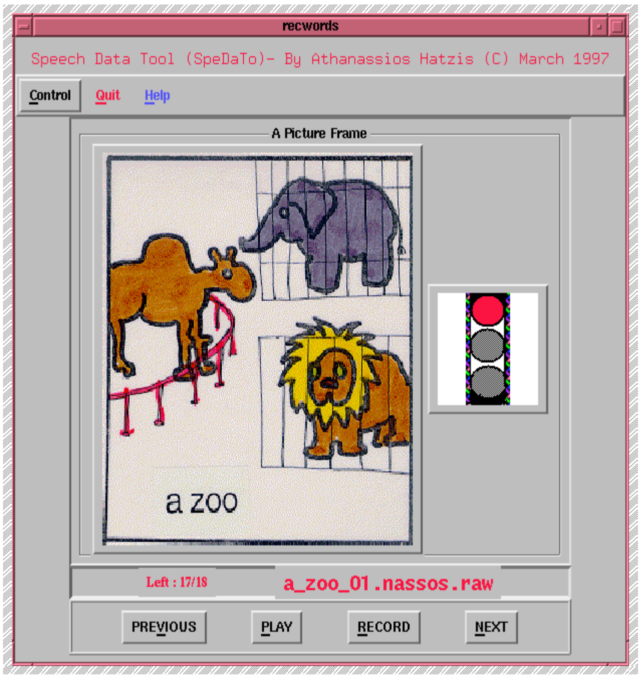
```
Option                                           Setting
-a       Output MELSPEC (mel spectrum)           Off
-b       Output FBANK (mel filter bank)          Off
-c       Output LPCEPSTRA (cepstral coef)        Off
-d       Append Delta Coef                       Off
-e       Append Log Energy                       On
-f T     Set frame period to T (msecs)           10.0
-g       Use power spectrum in MEL analysis      Magnitude
-h       Apply Hamming Window                    On
-i       Output IREFC (16 bit refl coef)         Off
-j       Subtract mean from coefficients         Off
-k f     Set PreEmphasis Coef to X               0.0
-l N     Set Cepstral Liftering Coef to N        0
-m       Output MFCC (mel freq cepstra)          On
-n N     Set number of parameters to N           8
-o       Replace log energy by 0'th MFCC         Off
-p N     Set parameter order to N                16
-q N     Set Delta window to -N..+N              Off
-r       Output LPREFC (refl coef)               Off
-s f     Scale Log Energy  or C_0 by f           0.1
-t       Suppress Log Energy Normalisation       Off
-u f     Low pass cut-off frequency (Hz)         Off
-v f     High pass cut-off frequency (Hz)        Off
-w T     Set window duration to T (msecs)        20
-x       Enable output compression               Off
-y f     Set silence floor (dBs)                 50.0
-z       Null energy (must have -d -e set)       Off
-F fmt   Set src file format to fmt              Off
-M Max   Set max of log energy for one frame     25.000
-O fmt   Set output file format to fmt           Off
-S Min   Set min log mean eng of silence         17.000
-T N     Set trace level to N                    0
```

Example : cat check.raw | HCodeRT -m -h -n 8 -e -p 16 -f 10 -w 20 -M 25 -S 17 -T 3

```
 HcodeRT               : (null) --> (null)
 Sample Kind           : MFCC_E
 Param Order           : 16
 Num Params            : 8
 Frame Period          : 10.00
 Frame Duration        : 20.00
 PreEmphasis           : 0.0000
 Hamming Window        : on
 Cep Liftering         : 0
 Src FileFormat        : undef
 Trace Level           : 3
FFT size = 512, fres = 0.044643
CFs
167.06   334.12   501.18   668.24   835.31  1002.37  1169.43  1336.49  1503.55  1670.61  1837.67
2004.73  2171.79  2338.85  2505.92  2672.98  2840.04
Wts
1.00   0.71   0.42   0.15   0.89   0.64   0.40   0.17   0.94   0.72   0.51   0.30   0.11   0.91
0.72   0.54   0.36   0.19   0.02   0.86   0.70   0.54   0.38   0.23 ...
Bins:
1.   0
2.   5. 1
3.   9. 2
4.   14. 3
5.   20. 4
6.   26. 5
7.   34. 6
8.   42. 7
9.   52. 8
10.  64. 9
11.  78. 10
12.  93. 11
13.  112. 12
14.  133. 13
15.  158. 14
16.  186. 15
17.  219. 16
```

```
0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
……..
MFCC Frame 200:
  1.  1.017141e+00   2. -6.472952e-01   3.  1.360187e+00   4. -5.907298e-01   5.  4.395519e-01   6. -
5.481228e-01   7.  4.649838e-01   8.  9.122248e-01   E.  1.863527e+01
MFCC Frame 201:
  1.  1.472647e+00   2.  2.958513e-01   3.  1.688668e+00   4.  3.002978e-01   5.  6.509624e-01   6. -
1.238359e+00   7.  1.378945e+00   8.  7.020164e-01   E.  2.037361e+01
MFCC Frame 202:
  1.  3.367644e+00   2. -4.813651e-01   3.  1.479180e+00   4. -3.655550e-01   5.  3.734667e-01   6. -
4.824230e-01   7.  3.098201e-01   8.  8.893427e-01   E.  2.271195e+01
MFCC Frame 203:
  1.  4.399112e+00   2. -1.570143e+00   3.  1.917195e+00   4. -1.011976e+00   5.  2.941042e-01   6. -
7.420002e-01   7.  2.346201e-01   8.  1.335419e+00   E.  2.349044e+01
………..
Normalising Log Energy
0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
………..
Frame :    200, Min = 17.000 (0.000), Val = 18.635 ( 0.364), Max = 25.000 (1.000)
1.017141 -0.647295 1.360187 -0.590730 0.439552 -0.548123 0.464984 0.912225 0.363527
Frame :    201, Min = 17.000 (0.000), Val = 20.374 ( 0.537), Max = 25.000 (1.000)
1.472647 0.295851 1.688668 0.300298 0.650962 -1.238359 1.378945 0.702016 0.537361
Frame :    202, Min = 17.000 (0.000), Val = 22.712 ( 0.771), Max = 25.000 (1.000)
3.367644 -0.481365 1.479180 -0.365555 0.373467 -0.482423 0.309820 0.889343 0.771195
Frame :    203, Min = 17.000 (0.000), Val = 23.490 ( 0.849), Max = 25.000 (1.000)
4.399112 -1.570143 1.917195 -1.011976 0.294104 -0.742000 0.234620 1.335419 0.849044
………..
```

C. Program SpeDaTo (Speech Data Tool)

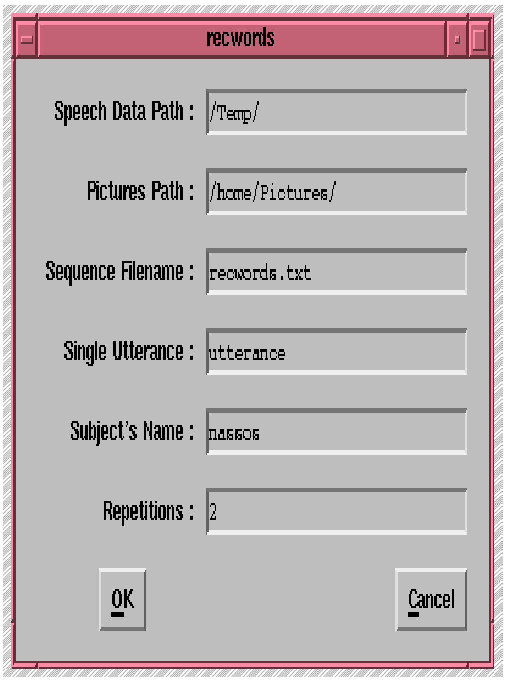


Speech Data Tool (SpeDaTo)- By Athanassios Hatzis (C) March 1997

Left : 17/18

a_zoo_01.nassos.raw

Figure Appendices-31 *Recording a sequence of utterances with SpeDaTo.*



Speech Data Path : /Temp/

Pictures Path : /home/Pictures/

Sequence Filename : recwords.txt

Single Utterance : utterance

Subject's Name : nassos

Repetitions : 2

Figure Appendices-32 *A dialogue window with several settings.*

Spedato is a graphics program for controlled recording/playback of a sequence of utterances. It has four buttons, similar to those on a cassette player (**Figure Appendices-31**). The *Record* button records the utterance displayed both in text and graphics format, *Play* repeats the last utterance recorded or plays back some other utterance in the sequence. *Previous, Next,* buttons locate the previous and next utterances in the sequence. During recording, the traffic light signals the start and end of the time duration.

A dialogue window of the program for several settings is also available (**Figure Appendices-32**). The user can change the location of the pictures, *Pictures Path*, the directory in which to store the utterances, *Speech Data Path*, and the file with the utterances to record, *Sequence Filename*. The last is an *ASCII* text file with all the utterances to record. The *Repetitions* field indicates how many times to repeat an utterance in a randomised sequence. The name of the recorded files is a concatenation of the utterance name and the *Subject's Name* field, together with a numeral at the end to distinguish between the repeated utterances.

D. Script autolyre_segment

Example 1 : autolyre_segment

```
Manual segmentation of files
Enter Speaker Directory : Dan
Processing file /home/nassos/Experiments/Frics_ISO/Dan/a_iso_01.dan.raw
 Attention ! /home/nassos/Experiments/Frics_ISO/Dan/a_iso_01.dan.lola is not written.

Processing file /home/nassos/Experiments/Frics_ISO/Dan/dh_iso_01.dan.raw
 Attention ! /home/nassos/Experiments/Frics_ISO/Dan/dh_iso_01.dan.lola is not written.

Processing file /home/nassos/Experiments/Frics_ISO/Dan/e_iso_01.dan.raw
 /home/nassos/Experiments/Frics_ISO/Dan/e_iso_01.dan.lola is successfully written.
```

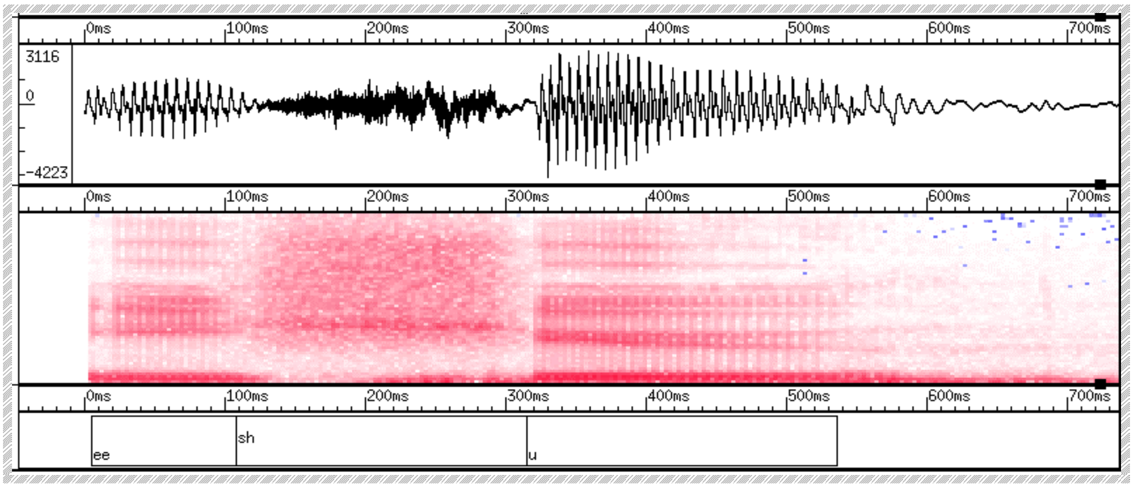


Figure Appendices-33 *The labelling of an utterance with the auto_lyre window. See also example 1 of ascii2dat for the labelled cepstral vectors.*

We usually run the *script* inside a directory where the utterances of each subject appear in separate subdirectories. The program asks for the name of a subdirectory and starts reading in alphabetical sequence the utterances files. Once a file is read it executes an *auto_lyre* window, **[22]**, split into three sections, time waveform, frequency spectrum, and labels (**Figure Appendices-33**). If the file is already labelled and it is not modified, an attention message is written on the terminal window; otherwise it is successfully written (Example 1). Labelling is done in *lyre* fashion, **[22]**, and a name is automatically given to the labels file according to the utterance name.

Example 2 : autolyre_segment e_iso_01.dan.raw

```
Manual segmentation of files
Enter Speaker Directory : Dan
Skipping /home/nassos/Experiments/Frics_ISO/Dan/a_iso_01.dan.raw
Skipping /home/nassos/Experiments/Frics_ISO/Dan/dh_iso_01.dan.raw
Skipping /home/nassos/Experiments/Frics_ISO/Dan/e_iso_01.dan.raw
Processing file /home/nassos/Experiments/Frics_ISO/Dan/e_iso_01.dan.raw
```

To inspect labelling of a certain utterance we have to pass the name of the file as an argument to *autolyre_segment*. The program automatically skips all the other files in the directory of the subject and displays the *auto_lyre* window of the utterance (**Example 2**).

E. The ascii2dat program

The command *ascii2dat* converts *ASCII* cepstral vectors to labelled or unlabeled cepstral data files using the label files of the *raw* format speech files.

```
usage: ascii2dat  [-m MODE] [-s FILE] [-i FILE(S)] [-o FILE] [-d Level] [-h help]
 Option         Represents                        Default
 -----------    ------------------------------    -------
 -m MODE        Mode (LAB,UNL,AVG,THR,CLA)         AVG
 -s FILE        Map stats filename
 -i FILE(S)     Input filename(s) *.ascii
 -o FILE        Output filename                   stdout
 -d Level       Debug Level                       0
 -h             Command usage
```

Example 1 : ascii2dat –m LAB –i “eeshu_12.dave.ascii” –d 3

```
eeshu_12.dave.Lola : eeshu_12.dave.ascii
Mode LAB : File STDOUT successfully opened
Processing 1 data files ...
eeshu_12.dave.ascii
Vectors : Read 75
Vectors=75, Dimensionality=9, Frequency=16000
[    5, 108]msec - [    0,  11]vec ee
[ 108, 315]msec - [  11,  32]vec sh
[ 315, 536]msec - [  32,  54]vec u
  0 : 2.601208 1.848445 5.278749 -0.069124 -0.594287 0.304902 -0.581969 -0.354075 0.327038 ee
  1 : 1.827517 1.965641 4.702361 0.181175 -0.351829 0.537448 -0.513559 -0.181587 0.335514 ee
  2 : 0.465356 1.986670 5.191573 0.053790 -0.666350 0.509158 -0.547015 -0.166756 0.358667 ee
  3 : -0.221527 2.044041 5.355179 0.160379 -0.588712 0.656318 -0.376014 -0.305062 0.366122 ee
  4 : -0.502300 2.153846 5.581469 0.409833 -0.777862 0.740367 -0.580205 -0.182701 0.385125 ee
  5 : -0.601865 2.187132 5.574967 0.314808 -0.599351 0.814608 -0.513232 -0.304006 0.376327 ee
  6 : -0.839445 2.362574 5.835666 0.409674 -0.781828 0.815032 -0.636877 -0.163109 0.373939 ee
  7 : -0.994084 2.461362 5.582992 0.302719 -0.551748 0.799168 -0.565024 -0.179415 0.386099 ee
  8 : -0.475936 2.497118 5.453719 0.497752 -0.510568 0.784959 -0.465449 -0.150233 0.362966 ee
  9 : 0.259517 2.403645 4.944698 0.591573 -0.317390 0.553171 -0.061589 -0.440906 0.327454 ee
 10 : 0.006810 2.696139 4.869291 0.762781 0.031101 0.624185 -0.007452 -1.011352 0.291290 ee
 11 : -1.027473 1.695645 4.003640 1.213486 0.510016 0.498218 0.446764 -0.403094 0.211513 sh
 12 : -3.772020 1.313504 3.991413 1.263979 0.693987 0.313739 0.056227 -0.573122 0.200000 sh
 13 : -5.612535 0.091867 2.527767 0.953684 0.309783 -0.049778 0.233063 -0.370903 0.200000 sh
 14 : -6.254655 0.079033 2.762489 0.912366 -0.215213 0.129931 0.217324 -0.025041 0.200000 sh
 15 : -6.723203 -0.026882 2.784742 0.467556 -0.693860 0.216965 0.446552 -0.040034 0.222341 sh
 16 : -6.037508 0.283043 2.273600 0.632347 -0.459441 -0.099563 0.129124 -0.334876 0.242984 sh
 17 : -6.598216 0.023288 2.653153 0.753939 -0.728892 0.123713 0.314074 -0.611256 0.256108 sh
 18 : -6.349419 0.351322 2.479064 0.980544 -0.400418 -0.335029 0.019169 -0.813782 0.243591 sh
 19 : -6.756034 0.383383 3.090905 0.561206 -0.628295 0.318864 0.301711 -0.544822 0.292696 sh
 20 : -5.850071 0.889513 3.244575 1.359824 0.029076 0.782921 0.770141 0.048278 0.310381 sh
 21 : -5.003609 1.172793 3.431507 1.648387 -0.188173 0.595608 0.983675 0.059999 0.307759 sh
```

```
22 : -5.099175 1.201758 3.703577 1.577399 -0.111449 0.689323 0.649961 -0.298870 0.322266 sh
23 : -3.976505 2.181152 3.303103 1.537133 -0.113010 0.417200 0.934990 -0.473997 0.336615 sh
24 : -4.467279 2.300067 3.701258 1.735038 -0.429141 0.646070 1.012492 -0.261159 0.374662 sh
25 : -4.203957 2.342385 4.469809 2.218307 -0.004949 0.400003 0.621839 -0.803755 0.364332 sh
26 : -3.901726 1.811729 3.311614 1.195320 -0.216266 0.460376 0.505351 -0.525037 0.301695 sh
27 : -5.000248 1.094281 3.387862 1.540048 -0.244195 0.553951 0.773943 -0.490070 0.286893 sh
28 : -2.966122 2.001273 3.864588 1.181488 -0.846135 0.092574 0.532235 -0.194155 0.287203 sh
29 : -2.456087 1.698118 3.428657 1.537426 -0.384727 0.224494 0.430533 -0.209811 0.242750 sh
30 : -1.373957 1.666972 3.879581 1.484751 -0.228423 0.027391 0.636684 -0.030437 0.200000 sh
31 : 1.559602 1.205966 3.479079 1.246480 -0.547550 -0.105621 0.019837 0.077408 0.368974 sh
32 : 0.262471 0.925295 4.609863 1.436576 -0.920747 -0.377040 -0.606489 0.330764 0.496268 u
33 : 0.661798 1.052661 4.539104 1.552306 -0.975562 -0.490517 -0.410178 0.286820 0.506230 u
34 : 0.558339 0.986219 4.229784 1.483989 -0.755364 -0.652266 -0.532199 0.280384 0.508756 u
35 : 0.593865 1.039887 4.271118 1.744998 -0.737765 -0.755623 -0.594025 0.319262 0.516348 u
36 : 0.507062 0.980602 3.940789 1.701021 -0.753842 -0.814821 -0.566644 0.402893 0.509951 u
37 : 0.435015 1.030602 3.961067 1.903116 -0.692632 -0.989986 -0.680397 0.479039 0.528627 u
38 : 0.568414 1.057094 3.567537 1.910846 -0.592722 -1.030088 -0.715168 0.595460 0.510685 u
39 : 0.800177 1.059057 3.456652 1.882292 -0.594649 -0.920956 -0.693600 0.536264 0.499176 u
40 : 1.140774 0.792011 3.580455 1.752960 -0.370860 -0.932038 -0.636737 0.591414 0.499455 u
41 : 1.603978 0.724639 3.689668 1.759056 -0.178725 -1.132588 -0.721173 0.538277 0.468857 u
42 : 2.125359 0.360948 3.847887 1.672897 -0.108615 -1.062284 -0.763471 0.600901 0.440220 u
43 : 2.525441 0.242316 3.920366 1.499015 0.099451 -1.243945 -0.787302 0.532878 0.440701 u
44 : 2.436847 0.541400 3.704532 1.667758 0.079073 -1.334263 -0.656279 0.518086 0.441750 u
45 : 3.248994 0.356730 3.609567 1.747773 0.002442 -0.972678 -1.021711 0.593090 0.438889 u
46 : 3.145660 0.574071 3.666608 1.582135 0.396053 -1.318942 -0.766991 0.313506 0.433261 u
47 : 3.137137 0.828668 3.498055 1.586763 0.539933 -1.316442 -0.679714 0.268654 0.423299 u
48 : 3.160736 0.754486 3.626359 1.425120 0.799420 -1.322108 -0.638092 0.126121 0.411741 u
49 : 3.438072 0.736622 3.639085 1.521407 0.613119 -1.067750 -0.823911 0.135770 0.399853 u
50 : 3.314217 1.210440 3.520207 1.368688 0.537921 -1.031205 -0.979431 0.386065 0.383374 u
51 : 3.161460 1.652327 3.163586 1.281505 0.685139 -1.005491 -0.803832 0.171450 0.367603 u
52 : 3.122947 2.241255 3.995424 1.514513 0.535654 -1.025462 -0.854424 0.283238 0.354870 u
53 : 3.570452 2.492251 2.962834 1.050390 0.403136 -0.718802 -0.698923 0.219716 0.332760 u
54 : 3.902359 2.438536 2.706873 1.221048 0.580450 -1.429171 -0.729973 -0.241081 0.337060 UNL
55 : 4.632211 2.403535 2.609528 1.905362 0.869663 -1.029898 -0.339644 -0.268884 0.333619 UNL
56 : 4.137850 2.740955 2.512139 1.610046 0.805411 -0.617880 -0.453487 -0.289393 0.337190 UNL
57 : 4.180078 3.160648 2.940138 1.668134 0.734168 -0.758795 0.051843 -0.186180 0.328904 UNL
58 : 2.862482 2.345745 2.223960 1.498976 0.999950 0.400568 0.405072 -0.110997 0.263492 UNL
59 : 2.857639 2.086650 2.231047 1.472915 0.815072 0.003441 0.347545 0.117141 0.233469 UNL
60 : 3.163906 1.958627 2.617233 1.321740 1.294728 0.024697 0.349271 -0.510298 0.206363 UNL
61 : 2.432845 1.478950 1.855077 1.290968 1.245066 -0.259209 -0.186349 0.084536 0.200000 UNL
62 : 2.785562 1.302893 2.198523 1.556250 1.314111 0.265838 0.272097 0.268847 0.200000 UNL
63 : 2.852801 1.121205 2.026649 1.938585 1.204839 -0.102203 0.032026 0.192946 0.200000 UNL
64 : 2.920081 1.174154 2.331027 1.327818 0.835663 -0.135645 0.237686 0.504731 0.200000 UNL
65 : 3.031117 1.578867 1.914202 1.309116 0.405905 0.068283 0.168246 -0.021717 0.200000 UNL
66 : 3.170730 1.656205 2.028225 1.631177 0.847520 -0.152955 0.206163 0.429887 0.200000 UNL
67 : 2.885308 1.825986 2.266699 1.649892 1.085282 0.128404 0.262244 0.357668 0.200000 UNL
68 : 2.612515 1.973781 2.410541 0.944314 1.161224 0.080509 -0.062766 0.677283 0.200000 UNL
69 : 2.715547 1.860125 2.101890 1.415090 0.771599 0.189058 0.255975 0.311212 0.200000 UNL
70 : 2.346842 0.739875 1.774601 1.495423 0.849239 0.032945 -0.186467 0.416410 0.200000 UNL
71 : 2.175344 1.034318 1.742210 1.507660 1.044170 0.497245 0.494191 0.376231 0.200000 UNL
72 : 2.967418 1.523934 2.192898 2.000451 1.465304 0.193863 -0.037284 0.108794 0.200000 UNL
73 : 2.143391 1.475535 1.642421 1.323051 1.141331 0.559220 0.482104 0.466565 0.200000 UNL
74 : 2.338980 0.824349 2.429267 2.130122 1.649403 0.202918 0.549420 0.529695 0.200000 UNL
```

Example 2 : ascii2dat -m AVG -i "*.*.ascii" -d 1 -o check

```
Processing 192 data files ...
~~~~~~~~~~  Vectors Read from ascii files ~~~~~~~~~~~~
ee : 2939 vectors
sh : 1152 vectors
  u : 4466 vectors
  s : 1133 vectors
 zh :  950 vectors
   z :  927 vectors
-----------------------------------------------------
Total number of vectors read                   :  13749
Total number of labelled vectors               :  11567
Average number of vectors between classes  :    1927
________________________________________________________
                    Create a dataset
*****************************************************
Enter the number of vectors per category : 800
*****************************************************
Do you want to create another data set from the remainder ? y
Enter the number of vectors per category : 200
_________________________________________________________
Selecting vectors for    ee category = 2939 vectors... Selected  800 vectors and  200 vectors
Selecting vectors for    sh category = 1152 vectors... Selected  800 vectors and  200 vectors
Selecting vectors for     u category  = 4466 vectors... Selected  800 vectors and  200 vectors
Selecting vectors for     s category  = 1133 vectors... Selected  800 vectors and  200 vectors
Selecting vectors for    zh category =  950  vectors... Selected  800 vectors and  150 vectors
Selecting vectors for     z category  =  927  vectors... Selected  800 vectors and  127 vectors
_________________________________________________________
Total number of vectors written for 1st set  :   4800
Total number of vectors written for 2nd set :  1077
________________________________________________________
```

Example 3 : ascii2dat –m CLA –s freexu.ldat –i “Words/Data/*.ascii” –d 1

```
Loading feature description file from features.nam
@@@@@@@@@@@@@@@@@@   O L T  - M A P    @@@@@@@@@@@@@@@@@@
   Name                        :       freexu.ldat
   Nr. of classes              :                 6
   Nr. of all samples          :              1077
   Nr. of all     features     :                 9
   Nr. of selected features    :                 9
   @@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
CLASS                NAME  Nr. of samples
 Nr.  1 :              ee     200
 Nr.  2 :               s     200
 Nr.  3 :              sh     200
 Nr.  4 :               u     200
 Nr.  5 :               z     127
 Nr.  6 :              zh     150
-------------------------------------------------------------
--- FEATURES OF DATA --- : 9
Parameter.   1 =   1       Name: mfcc_01
Parameter.   2 =   2       Name: mfcc_02
Parameter.   3 =   3       Name: mfcc_03
Parameter.   4 =   4       Name: mfcc_04
Parameter.   5 =   5       Name: mfcc_05
Parameter.   6 =   6       Name: mfcc_06
Parameter.   7 =   7       Name: mfcc_07
Parameter.   8 =   8       Name: mfcc_08
Parameter.   9 =   9       Name: norm_eng
-------------------------------------------------------------
Processing 192 data files ...
 ~~~~~~~~~~  Vectors Read from ascii files ~~~~~~~~~~~~
  ee     : 2939 vectors
  sh     : 1152 vectors
   u     : 4466 vectors
   s     : 1133 vectors
  zh     :   950 vectors
   z     :   927 vectors
Total number of vectors read               :  13749
Total number of labelled vectors           :  11567
Average number of vectors between classes  :    1927
```

Classification according to maximum a posteriori probability NN

~~~~~~~~~~~~~ Classification Results ~~~~~~~~~~~~~~~~~~

Correct Classified : 10900 --- Wrong Classified : 667

Error Rate = 5.77% --- **Accuracy 94.23%**

~~~~~~ C O N F U S I O N M A T R I X ~~~~~~

ROW: expected class --- COLUMN: detected class

# test samples: 11567

Relative	**ee**	**s**	**sh**	**u**	**z**	**zh**
ee	**96.33%**	0.51%	1.09%	0.24%	0.54%	1.29%
s	1.32%	**88.35%**	1.15%	0.71%	7.50%	0.97%
sh	1.39%	0.95%	**87.33%**	0.26%	0.09%	9.98%
u	0.20%	0.47%	0.22%	**97.54%**	0.43%	1.14%
z	1.19%	4.64%	0.65%	1.08%	**91.80%**	0.65%
zh	1.89%	0.53%	6.11%	0.21%	1.26%	**90.00%**

Classification according to a hyperquadratic discrimination function

~~~~~~~~~~~~~ Classification Results ~~~~~~~~~~~~~~~~~~

Correct Classified : 10663 --- Wrong Classified : 904

Error Rate = 7.82% --- Accuracy 92.18%

~~~~~~ C O N F U S I O N M A T R I X ~~~~~~

ROW: expected class --- COLUMN: detected class

# test samples: 11567

Relative	**ee**	**s**	**sh**	**u**	**z**	**zh**
ee	**95.24%**	0.68%	0.24%	0.41%	0.54%	2.89%
s	1.50%	**83.67%**	1.68%	1.85%	8.65%	2.65%
sh	2.69%	1.82%	**81.86%**	1.22%	0.69%	11.72%
u	0.34%	0.13%	0.02%	**98.52%**	0.38%	0.60%
z	1.51%	6.15%	0.00%	1.83%	**87.06%**	3.45%
zh	2.95%	0.95%	11.26%	3.37%	0.84%	**80.63%**

Results from running command in Example 3 for other methods and data sets

File name	*'iso2vo4frlva'*				
Description	Normal English male adults map with two vowels and four sibilant fricative targets and an LVQ codebook set.				
Subjects analysed	8 normal English male adults				
Data analysed	Vowels and consonants in isolation				
Classes	**Data set**	**Train set HDF**	**LVQ-6x100 HDF**	**Train set NN-9x16x6**	**LVQ-6x100 knn-1**
/i/	2042	99.71%	99.61%	99.85%	100.00%
/u/	2079	100.00%	100.00%	100.00%	100.00%
/s/	1883	100.00%	100.00%	100.00%	100.00%
/ʃ/	2049	99.89%	99.84%	99.95%	99.84%
/z/	2266	99.87%	99.87%	100.00%	100.00%
/ʒ/	2207	99.91%	99.86%	99.95%	99.64%
Total	**12526**	**99.90%**	**99.86%**	**99.96%**	**99.91%**

Table Appendices-2 *Comparison of classification results from isolated phonemes between four different methods. The best one is the connectionist method PPANN.*

File name	*'freeXu_lva'*				
Description	Normal English male adults map with two vowels and four sibilant fricative targets and an LVQ codebook set.				
Subjects analysed	6 normal English male adults – 8 repetitions each				
Data analysed	Vowels and consonants in i X u context, X is (s, ʃ, z, ʒ).				
Classes	**Data set**	**Train set HDF**	**LVQ-6x100 HDF**	**Train set NN-9x16x6**	**LVQ-6x100 knn-1**
/i/	2939	95.24%	91.56%	96.33%	96.94%
/u/	1152	98.52%	96.95%	97.54%	97.98%
/s/	4466	83.67%	81.64%	88.35%	82.79%
/ʃ/	1133	81.86%	81.68%	87.33%	84.72%
/z/	950	87.06%	87.81%	91.80%	94.07%
/ʒ/	927	80.63%	80.42%	90.00%	90.11%
Total	**11567**	**92.18%**	**90.47%**	**94.23%**	**93.95%**

Table Appendices-3 *Comparison of classification results from phones in ee_X_u context between four different methods. The best one is the connectionist method PPANN.*

Example 4 : ascii2dat –m THR -s iso2vo4frlvq.cod -i "ISO/Invalid/*.ascii" -d 5

```
Mode THR : File STDOUT successfully opened
Processing 72 data files ...
 ~~~~~~~~~~  Vectors Read from ascii files ~~~~~~~~~~~~
   a       : 1920 vectors
   dh      : 1748 vectors
   e       : 2007 vectors
   f       : 2072 vectors
   m       : 2182 vectors
   n       : 2281 vectors
   o       : 2033 vectors
   th      : 1849 vectors
   v       : 2020 vectors
---------------------------------------------------------------------------
Total number of vectors read                 :  21178
Total number of labelled vectors             :  18112
Average number of vectors between classes   :    2012

Classification according to a hyperquadratic discrimination function

..........
SAMPLE :   4.76   0.99   -0.74   -1.22    0.87    1.48    0.22   -0.19    0.62   a
MAP : -363.945/i/, -647.783/s/, -205.451/sh/, -165.325/u/, -203.511/z/, -202.689/zh/,
..........
SAMPLE :   1.43   2.43    0.48    0.61    0.55    0.25   -0.07   -0.66    0.41   dh
MAP : -66.4021/i/, -183.992/s/, -82.3243/sh/, -73.2533/u/, -9.96693/z/, -60.4238/zh/,
..........
SAMPLE :   2.99   1.14    2.01    0.64   -0.95   -0.44   -0.23    0.52    0.59   e
MAP : -37.9972/i/, -433.932/s/, -77.5772/sh/, -81.5989/u/, -52.6827/z/, -55.6512/zh/,
```

Threshold Loop [-20,10] step 0.5

```
........
*** Threshold = -2.000, Accepted =   1202, Rejected =  16910 ***
 -2.000   6.636  93.364 (Output for the gnuplot)
Category =        a, Threshold      = -2.000, Accepted =     0, Rejected =   1920
Category =       dh, Threshold      = -2.000, Accepted =    90, Rejected =   1658
Category =        e, Threshold      = -2.000, Accepted =     0, Rejected =   2007
Category =        f, Threshold      = -2.000, Accepted =     5, Rejected =   2067
Category =        m, Threshold      = -2.000, Accepted =   499, Rejected =   1683
Category =        n, Threshold      = -2.000, Accepted =   516, Rejected =   1765
Category =        o, Threshold      = -2.000, Accepted =     0, Rejected =   2033
Category =       th, Threshold      = -2.000, Accepted =    18, Rejected =   1831
Category =        v, Threshold      = -2.000, Accepted =    74, Rejected =   1946
.........
*** Threshold =  2.000, Accepted =    291, Rejected =  17821 ***
2.000   1.607  98.393 (Output for the gnuplot)
Category =        a, Threshold      =  2.000, Accepted =     0, Rejected =   1920
Category =       dh, Threshold      =  2.000, Accepted =    11, Rejected =   1737
Category =        e, Threshold      =  2.000, Accepted =     0, Rejected =   2007
Category =        f, Threshold      =  2.000, Accepted =     1, Rejected =   2071
Category =        m, Threshold      =  2.000, Accepted =   131, Rejected =   2051
Category =        n, Threshold      =  2.000, Accepted =   141, Rejected =   2140
Category =        o, Threshold      =  2.000, Accepted =     0, Rejected =   2033
Category =       th, Threshold      =  2.000, Accepted =     0, Rejected =   1849
Category =        v, Threshold      =  2.000, Accepted =     7, Rejected =   2013
.........
```

F. The sigmoid mean square artificial neural network (SMANN)

```
Usage : smnn    [-m MODE] [-i INPUT] [-h HIDDEN] [-o OUTPUT] [-l LRATE] [-c CYCLES]
                [-t FILE]  [-v FILE] [-w WEIGHTS] [-n NDfile] [-d DEBUG]
where:
         Option          Represents                        Default
         -----------     ------------------------------    -------
         -m MODE         Mode (TRN,TST,LAB,UNL)
         -i INPUT        Input layer units                  9
         -h HIDDEN       Hidden layer units                 8
         -o OUTPUT       Output layer units                 7
         -l LRATE        Learning rate                      0.001
         -c CYCLES       Number of training cycles         1000
         -t FILE         Training patterns file (SNNS)
         -v FILE         Testing patterns file (SNNS)
         -w WEIGHTS      Weights matrix file               weights
         -n NDfile       NDimensional vectors file
         -d DEBUG        Debug level                       0
```

Example 1 :   smnn –m TRN –i 9 –h 16 –o 2 –l 0.01 –c 50
–t train9D2D.pat –v test9D2D.pat –d 3

```
Training patterns file name : train9D2D.pat - Size :  12000 patterns
Testing  patterns file name  : test9D2D.pat - Size   :    3576 patterns
Weights matrix file name    : weights
Net structure 9 input x 16 hidden x 2 output - Weights :    194
Learning rate = 0.010000, Training cycles = 200
Minimum Mean Square Error = 0.000100
Cycle :    1 - Train MSSE :    0.0189 - Test MSSE :    0.0466
Cycle :    2 - Train MSSE :    0.0139 - Test MSSE :    0.0402
Cycle :    3 - Train MSSE :    0.0121 - Test MSSE :    0.0377
.........................................................................................
Cycle :  196 - Train MSSE :    0.0038 - Test MSSE :    0.0113
Cycle :  197 - Train MSSE :    0.0038 - Test MSSE :    0.0113
Cycle :  198 - Train MSSE :    0.0038 - Test MSSE :    0.0113
Cycle :  199 - Train MSSE :    0.0038 - Test MSSE :    0.0113
Cycle :  200 - Train MSSE :    0.0038 - Test MSSE :    0.0113
Training started   at : Tue May 25 21:29:28 1999
Training finished at : Tue May 25 23:04:31 1999
Duration of training : 5703 secs
```

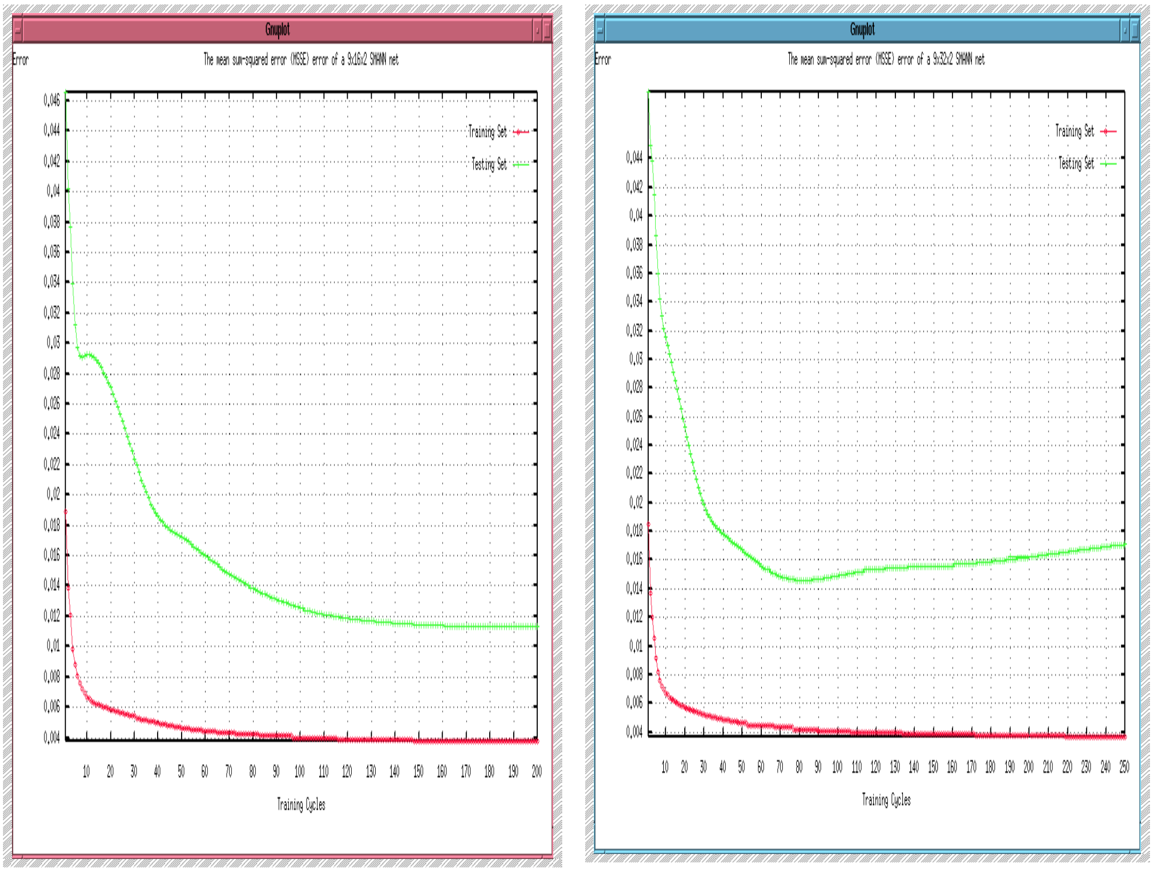


Figure Appendices-34 *The (MSSE) error of SMANN for a training and testing set. On the left we see the previous run-time output example plotted and on the right we see a case of overtraining by increasing the number of hidden units.*

Example 2 : smnn -m TST -i 9 -h 16 -o 2 -w wgts.9x16x2 -v test9D2D.pat

```
Testing  patterns file name : test9D2D.pat - Size :   3000 patterns

Weights matrix file name    : wgts.9x16x2

0.344 0.264

0.342 0.258

0.344 0.259

0.340 0.263

0.335 0.264

0.333 0.268

.....................

Test MSSE :    0.0039
```

G. The a posteriori probabilities artificial neural network (PPANN)

```
Usage: ppnn     [-m MODE] [-i INPUT] [-h HIDDEN] [-o OUTPUT] [-l LRATE] [-c CYCLES]
                [-t FILE] [-v FILE] [-w WEIGHTS] [-d DEBUG]
where:
        Option          Represents                      Default
        -----------     ------------------------------  -------
        -m MODE         Mode (TRAIN,TEST)
        -i INPUT        Input layer units               9
        -h HIDDEN       Hidden layer units              8
        -o OUTPUT       Output layer units        7
        -l LRATE        Learning rate                   0.001
        -c CYCLES       Number of training cycles       1000
        -t FILE         Training patterns file (SNNS)
        -v FILE         Testing patterns file (SNNS)
        -w WEIGHTS      Weights matrix file             weights
        -d DEBUG        Debug level                     0
```

Example 1 : ppnn –m TEST –i 9 –h 16 –o 6 –l 0.001 –c 350 –t trainCLASS.pat –v testCLASS.pat

```
Training patterns file name : trainCLASS.pat - Size :  12000 patterns
Testing  patterns file name : testCLASS.pat - Size   :   3000 patterns
Weights matrix file name    : weights
Coding scheme of patterns   : i 0 o 1 s 2 s1 3 s2 4 sh 5
Net structure 9 input x 16 hidden x 6 output - Weights :    262
Learning rate   = 0.001000, Training cycles = 350
Minimum entropy = 0.001000
Cycle :    1 - Train Entropy :    1.1064 - Test Entropy :    1.1082
Cycle :    2 - Train Entropy :    0.7850 - Test Entropy :    0.7914
Cycle :    3 - Train Entropy :    0.6225 - Test Entropy :    0.6337
Cycle :    4 - Train Entropy :    0.5496 - Test Entropy :    0.5656
..............................................................................................................
Cycle :  345 - Train Entropy :    0.1400 - Test Entropy :    0.1621
Cycle :  346 - Train Entropy :    0.1400 - Test Entropy :    0.1621
Cycle :  347 - Train Entropy :    0.1399 - Test Entropy :    0.1621
Cycle :  348 - Train Entropy :    0.1399 - Test Entropy :    0.1621
Cycle :  349 - Train Entropy :    0.1399 - Test Entropy :    0.1621
Cycle :  350 - Train Entropy :    0.1398 - Test Entropy :    0.1621
Training started   at  : Wed May  26 20:45:21 1999
Training finished at  : Thu  May  27 00:59:46 1999
Duration of training  : 15265 secs
```

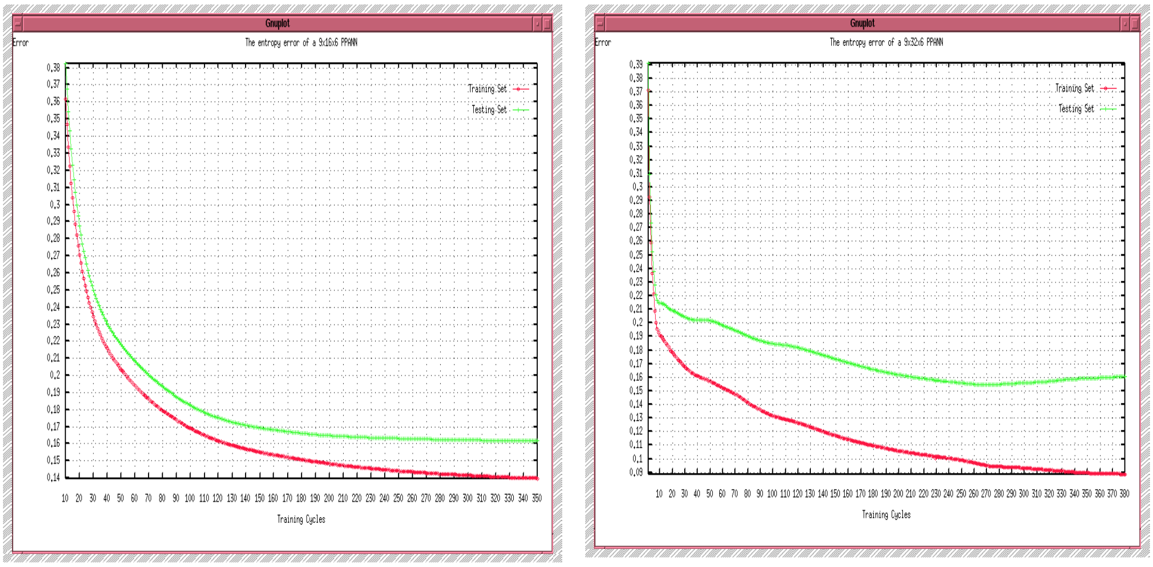


Figure Appendices-35 *The entropy error of PPANN for a training and testing set. On the left we see the previous run-time output example plotted and on the right we see a case of overtraining by increasing the number of hidden units.*

Example 2 : ppnn -m TEST -i 9 -h 32 -o 6 -w wgts.9x32x6 -v testCLASS.pat -d 3

```
Testing  patterns file name  : testCLASS.pat - Size :   3000 patterns
Weights matrix file name    : wgts.9x32x6
1.000 0.000 0.000 0.000 0.000 0.000 i
0.990 0.000 0.000 0.009 0.001 0.000 i
0.999 0.001 0.000 0.000 0.000 0.000 i
………………………………………..
0.000 1.000 0.000 0.000 0.000 0.000 o
0.001 0.999 0.000 0.000 0.000 0.000 o
0.011 0.989 0.000 0.000 0.000 0.000 o
………………………………………
0.000 0.000 0.859 0.141 0.000 0.000 s
0.000 0.000 1.000 0.000 0.000 0.000 s
0.000 0.000 0.235 0.319 0.446 0.000 s2
………………………………………..

0.000 0.000 0.000 1.000 0.000 0.000 s1
0.000 0.000 0.012 0.988 0.000 0.000 s1
0.000 0.000 0.739 0.261 0.000 0.000 s
…………………………………………
0.000 0.000 0.000 0.002 0.998 0.000 s2
0.000 0.000 0.009 0.039 0.953 0.000 s2
0.000 0.000 0.601 0.340 0.059 0.000 s
…………………………………………
0.000 0.000 0.000 0.000 0.000 1.000 sh
0.000 0.000 0.000 0.109 0.000 0.891 sh
0.000 0.000 0.064 0.785 0.000 0.151 s1
Test Entropy :    0.1558
~~~~~~~~~~~~~ Classification Results ~~~~~~~~~~~~~~~~~
 Correct Classified : 2839 --- Wrong Classified : 161
 Error Rate =   5.37% --- Accuracy  94.63%
```

H. Creating a phonetic map

Commands for creating a phonetic map	
Mkdir Data	Create a directory to store the data files
lnk ../../Analysis/*.raw .	Link the required speech files from the analysis pool.
lnk ../../Analysis/*.Lola .	Link the labels of the speech files
raw2cepstral *.raw	Convert speech data to binary cepstral vectors
cepstral2ascii *.hmc	Convert binary cepstral vectors to ascii cepstral vectors
rm –f *.hmc	Remove all binary cepstral vectors
Ascii2dat –m AVG –i "*.ascii" –o train.ldat –d 2 Collects all the vectors from the speech data files and groups them under categories. Calculates the average number of vectors per category and randomly selects a number of vectors from the data to create the training and the testing files.	
mkdir NN	Create a directory to store files for training the ANNs
copy *.awk	Copy *awk* scripts and edit them appropriately
lnk ../Data/train.ldat . **lnk ../Data/test.ldat .**	Link the training and testing data set from the Data directory in the NN directory
ndxshuffle train.ldat nrsamples > trainrnd.ldat Randomises the training set	
gawk –f 1-of-n.awk > trainCLASS.pat [ENTER] **trainrnd.ldat nrsamples [CTRL+D]** **gawk –f 1-of-n.awk > testCLASS.pat [ENTER]** **test.ldat nrsamples [CTRL+D]**	Create the training and testing pattern files for use in the *PPANN* classifier. The pattern files follow the format of SNNS.
gawk –f ldat2pat.awk > train9D2D.pat [ENTER] **trainrnd.ldat nrsamples [CTRL+D]** **gawk –f ldat2pat.awk > test9D2D.pat [ENTER]** **test.ldat nrsamples [CTRL+D]**	Create the training and testing pattern files for use in the *SMANN* non-linear mapping function. The pattern files follow the format of SNNS.
ppnn –m TRAIN –i 9 –h 16 –o 6 –l 0.01 –c 100 –w wgts9x16x6 –t trainCLASS.pat –v testCLASS.pat –d 3 Sample command to train the PPANN classifier	
smnn –m TRN –i 9 –h 16 –o 2 –l 0.01 –c 100 –w wgts9x16x2 –t train9D2D.pat –v test9D2D.pat –d 3 Sample command to train the SMANN mapping function	
Eveninit –noc 600 –din train.ldat –cout lvq.ini –knn 5 **olvq1 –din train.ldat –cin lvq.ini –cout lvq.cod –rlen 24000** Commands to initialise and train a codebook vector set using the *LVQ* method of Kohonen.	
Create the files of the phonetic map with the appropriate names in the Maps directory. **Fname.ixhxo.smn, fname.ixhxo.ppn, fname.cod, fname.fts, fname.map**	

Table Appendices-4 *Unix commands and scripts to create a phonetic map*

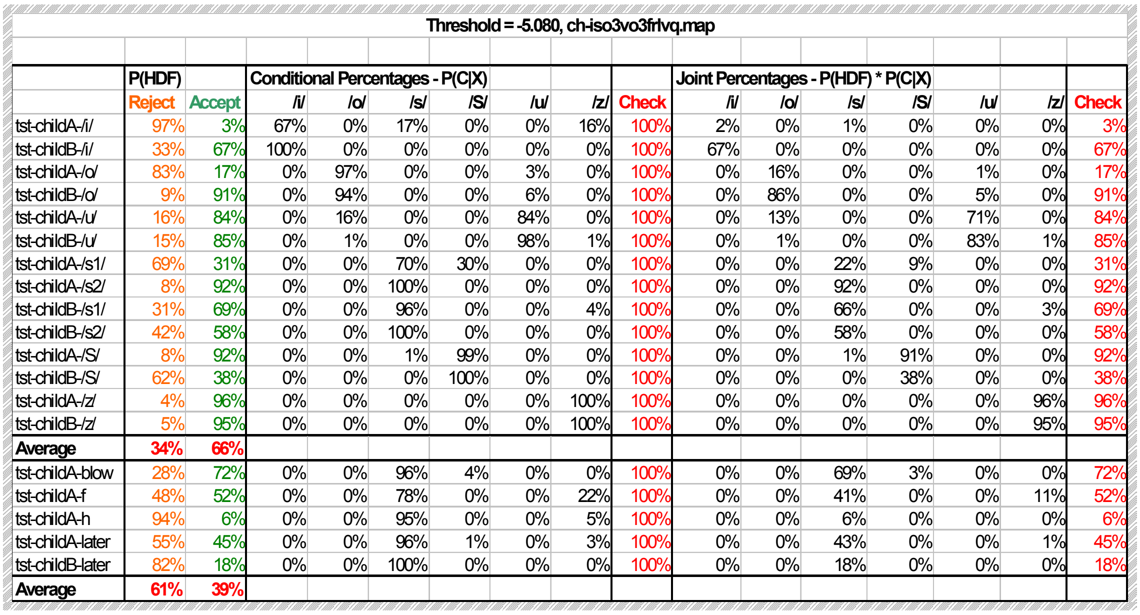
I. Recording pipeline of OLTK

Example :

```
rec -f /Temp/recbuffer.raw |
sildetect -d 0 -p 0 -m 12.320000 -s 13.630000 -e 17.370001 -a 0.200000 -b 0.060000  |
HCodeRT -m -h -n 8 –e -p 16 -f 10 -w 20 -M 25.000000 -S 17.000000
```

```
Threshold = -20.000000
Relax threshold 0.000000,  0.000000,  0.000000,  0.000000,  0.000000,  0.000000,
Animation cycles time        = 60 sec
Ellapsed time of animation   = 60.000000 sec
Animation Frequency          = 6000/60.000000 = 100 Hz
```

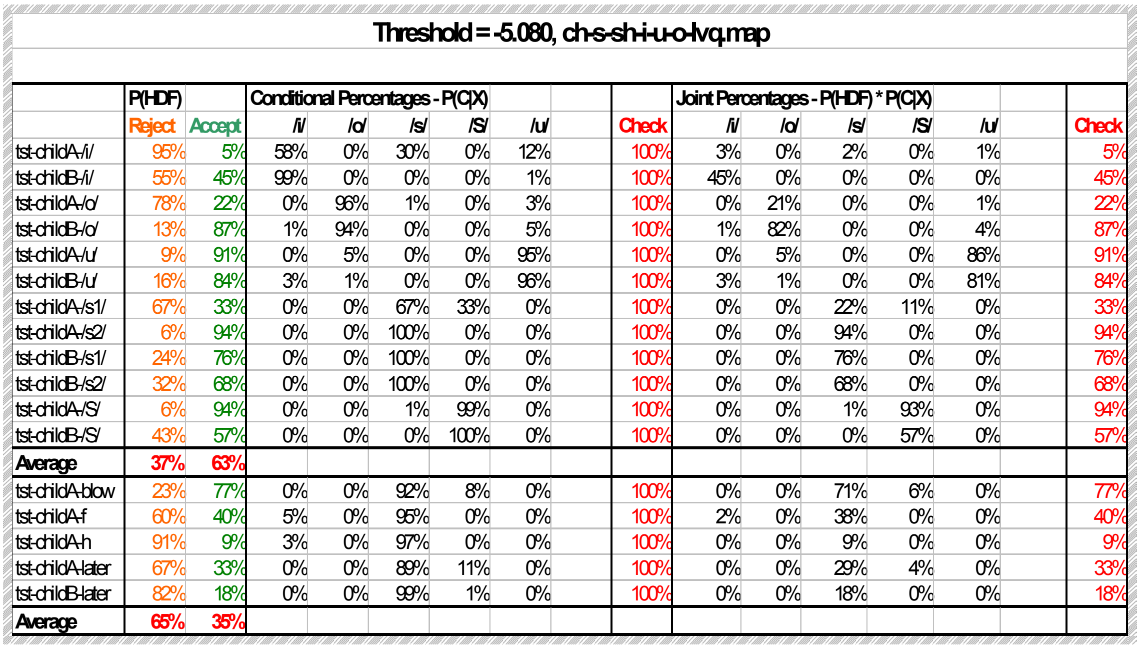
IV Independent Speakers Map Testing



Threshold = -5.080, ch-iso3vo3frlvq.map

	P(HDF)		Conditional Percentages - P(C|X)							Joint Percentages - P(HDF) * P(C|X)						
	Reject	Accept	/i/	/o/	/s/	/S/	/u/	/z/	Check	/i/	/o/	/s/	/S/	/u/	/z/	Check
tst-childA-/i/	97%	3%	67%	0%	17%	0%	0%	16%	100%	2%	0%	1%	0%	0%	0%	3%
tst-childB-/i/	33%	67%	100%	0%	0%	0%	0%	0%	100%	67%	0%	0%	0%	0%	0%	67%
tst-childA-/o/	83%	17%	0%	97%	0%	0%	3%	0%	100%	0%	16%	0%	0%	1%	0%	17%
tst-childB-/o/	9%	91%	0%	94%	0%	0%	6%	0%	100%	0%	86%	0%	0%	5%	0%	91%
tst-childA-/u/	16%	84%	0%	16%	0%	0%	84%	0%	100%	0%	13%	0%	0%	71%	0%	84%
tst-childB-/u/	15%	85%	0%	1%	0%	0%	98%	1%	100%	0%	1%	0%	0%	83%	1%	85%
tst-childA-/s1/	69%	31%	0%	0%	70%	30%	0%	0%	100%	0%	0%	22%	9%	0%	0%	31%
tst-childA-/s2/	8%	92%	0%	0%	100%	0%	0%	0%	100%	0%	0%	92%	0%	0%	0%	92%
tst-childB-/s1/	31%	69%	0%	0%	96%	0%	0%	4%	100%	0%	0%	66%	0%	0%	3%	69%
tst-childB-/s2/	42%	58%	0%	0%	100%	0%	0%	0%	100%	0%	0%	58%	0%	0%	0%	58%
tst-childA-/S/	8%	92%	0%	0%	1%	99%	0%	0%	100%	0%	0%	1%	91%	0%	0%	92%
tst-childB-/S/	62%	38%	0%	0%	0%	100%	0%	0%	100%	0%	0%	0%	38%	0%	0%	38%
tst-childA-/z/	4%	96%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	0%	96%	96%
tst-childB-/z/	5%	95%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	0%	95%	95%
Average	34%	66%														
tst-childA-blow	28%	72%	0%	0%	96%	4%	0%	0%	100%	0%	0%	69%	3%	0%	0%	72%
tst-childA-f	48%	52%	0%	0%	78%	0%	0%	22%	100%	0%	0%	41%	0%	0%	11%	52%
tst-childA-h	94%	6%	0%	0%	95%	0%	0%	5%	100%	0%	0%	6%	0%	0%	0%	6%
tst-childA-later	55%	45%	0%	0%	96%	1%	0%	3%	100%	0%	0%	43%	0%	0%	1%	45%
tst-childB-later	82%	18%	0%	0%	100%	0%	0%	0%	100%	0%	0%	18%	0%	0%	0%	18%
Average	61%	39%														

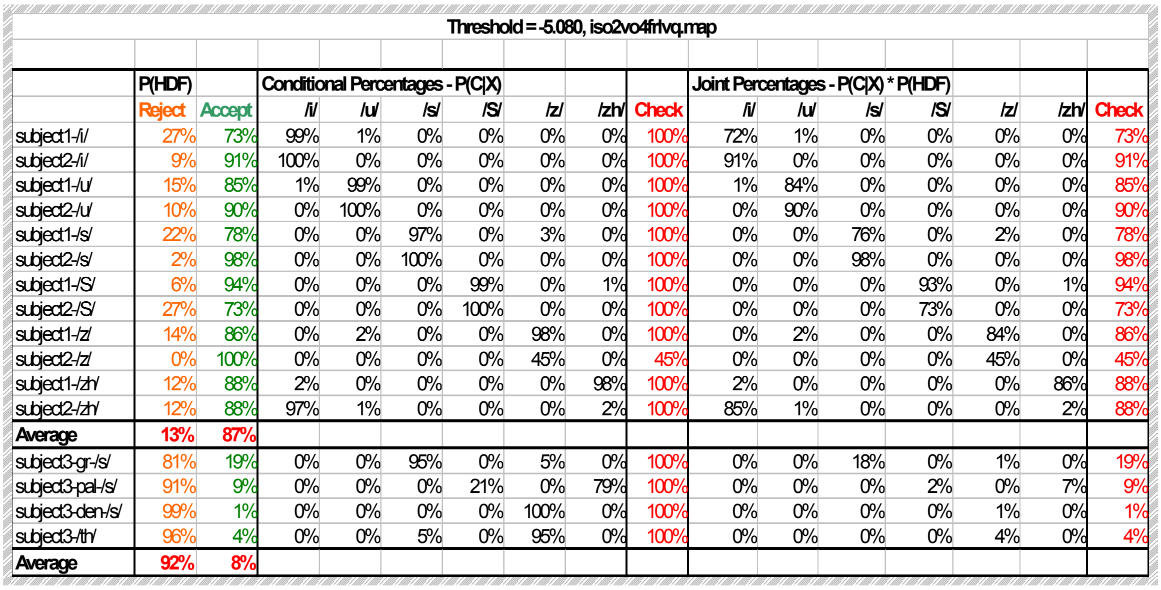
Table Appendices-5 : *Independent speakers testing of 'ch-iso3vo3frlvq' map*



Threshold = -5.080, ch-s-sh-i-u-o-lvq.map

	P(HDF)		Conditional Percentages - P(C|X)							Joint Percentages - P(HDF) * P(C|X)						
	Reject	Accept	/i/	/o/	/s/	/S/	/u/		Check	/i/	/o/	/s/	/S/	/u/		Check
tst-childA-/i/	95%	5%	58%	0%	30%	0%	12%		100%	3%	0%	2%	0%	1%		5%
tst-childB-/i/	55%	45%	99%	0%	0%	0%	1%		100%	45%	0%	0%	0%	0%		45%
tst-childA-/o/	78%	22%	0%	96%	1%	0%	3%		100%	0%	21%	0%	0%	1%		22%
tst-childB-/o/	13%	87%	1%	94%	0%	0%	5%		100%	1%	82%	0%	0%	4%		87%
tst-childA-/u/	9%	91%	0%	5%	0%	0%	95%		100%	0%	5%	0%	0%	86%		91%
tst-childB-/u/	16%	84%	3%	1%	0%	0%	96%		100%	3%	1%	0%	0%	81%		84%
tst-childA-/s1/	67%	33%	0%	0%	67%	33%	0%		100%	0%	0%	22%	11%	0%		33%
tst-childA-/s2/	6%	94%	0%	0%	100%	0%	0%		100%	0%	0%	94%	0%	0%		94%
tst-childB-/s1/	24%	76%	0%	0%	100%	0%	0%		100%	0%	0%	76%	0%	0%		76%
tst-childB-/s2/	32%	68%	0%	0%	100%	0%	0%		100%	0%	0%	68%	0%	0%		68%
tst-childA-/S/	6%	94%	0%	0%	1%	99%	0%		100%	0%	0%	1%	93%	0%		94%
tst-childB-/S/	43%	57%	0%	0%	0%	100%	0%		100%	0%	0%	0%	57%	0%		57%
Average	37%	63%														
tst-childA-blow	23%	77%	0%	0%	92%	8%	0%		100%	0%	0%	71%	6%	0%		77%
tst-childA-f	60%	40%	5%	0%	95%	0%	0%		100%	2%	0%	38%	0%	0%		40%
tst-childA-h	91%	9%	3%	0%	97%	0%	0%		100%	0%	0%	9%	0%	0%		9%
tst-childA-later	67%	33%	0%	0%	89%	11%	0%		100%	0%	0%	29%	4%	0%		33%
tst-childB-later	82%	18%	0%	0%	99%	1%	0%		100%	0%	0%	18%	0%	0%		18%
Average	65%	35%														

Table Appendices-6 : *Independent speakers testing of 'ch-s-sh-i-u-o-lvq' map*

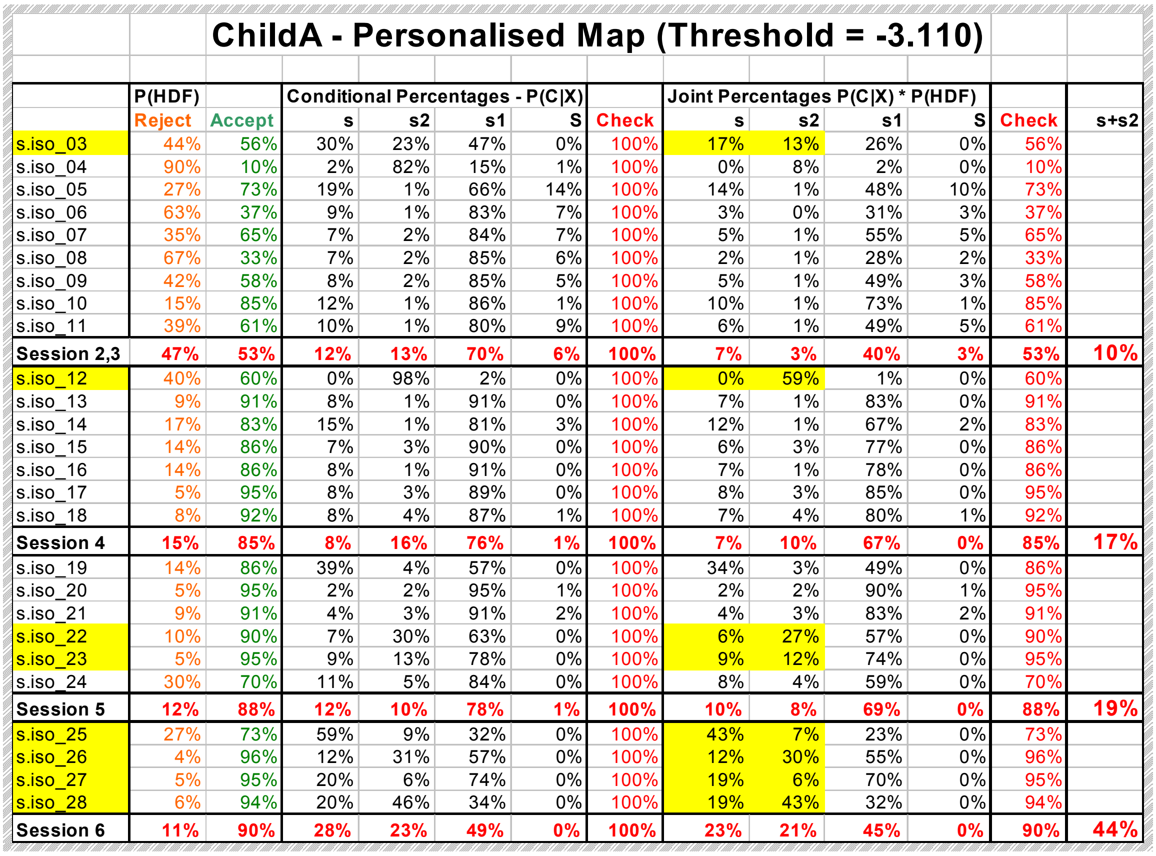


Threshold = -5.080, iso2vo4frlvq.map

	P(HDF)		Conditional Percentages - P(C|X)							Joint Percentages - P(C|X) * P(HDF)						
	Reject	Accept	/i/	/u/	/s/	/S/	/z/	/zh/	Check	/i/	/u/	/s/	/S/	/z/	/zh/	Check
subject1-/i/	27%	73%	99%	1%	0%	0%	0%	0%	100%	72%	1%	0%	0%	0%	0%	73%
subject2-/i/	9%	91%	100%	0%	0%	0%	0%	0%	100%	91%	0%	0%	0%	0%	0%	91%
subject1-/u/	15%	85%	1%	99%	0%	0%	0%	0%	100%	1%	84%	0%	0%	0%	0%	85%
subject2-/u/	10%	90%	0%	100%	0%	0%	0%	0%	100%	0%	90%	0%	0%	0%	0%	90%
subject1-/s/	22%	78%	0%	0%	97%	0%	3%	0%	100%	0%	0%	76%	0%	2%	0%	78%
subject2-/s/	2%	98%	0%	0%	100%	0%	0%	0%	100%	0%	0%	98%	0%	0%	0%	98%
subject1-/S/	6%	94%	0%	0%	0%	99%	0%	1%	100%	0%	0%	0%	93%	0%	1%	94%
subject2-/S/	27%	73%	0%	0%	0%	100%	0%	0%	100%	0%	0%	0%	73%	0%	0%	73%
subject1-/z/	14%	86%	0%	2%	0%	0%	98%	0%	100%	0%	2%	0%	0%	84%	0%	86%
subject2-/z/	0%	100%	0%	0%	0%	0%	45%	0%	45%	0%	0%	0%	0%	45%	0%	45%
subject1-/zh/	12%	88%	2%	0%	0%	0%	0%	98%	100%	2%	0%	0%	0%	0%	86%	88%
subject2-/zh/	12%	88%	97%	1%	0%	0%	0%	2%	100%	85%	1%	0%	0%	0%	2%	88%
Average	13%	87%														
subject3-gr-/s/	81%	19%	0%	0%	95%	0%	5%	0%	100%	0%	0%	18%	0%	1%	0%	19%
subject3-pal-/s/	91%	9%	0%	0%	0%	21%	0%	79%	100%	0%	0%	0%	2%	0%	7%	9%
subject3-den-/s/	99%	1%	0%	0%	0%	0%	100%	0%	100%	0%	0%	0%	0%	1%	0%	1%
subject3-/th/	96%	4%	0%	0%	5%	0%	95%	0%	100%	0%	0%	0%	0%	4%	0%	4%
Average	92%	8%														

Table Appendices-7 : *Independent speakers testing of 'iso2vo4frlvq' map*

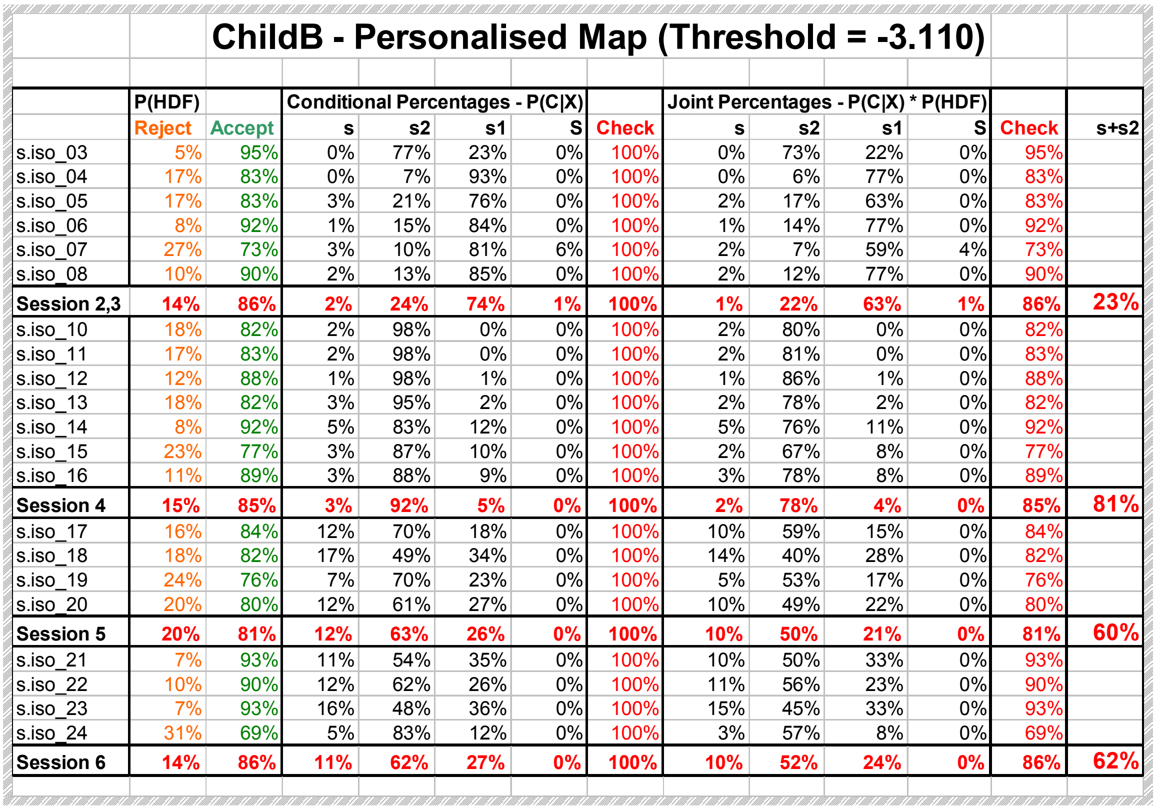
V Speech Therapy Results



ChildA - Personalised Map (Threshold = -3.110)

	P(HDF)		Conditional Percentages - P(C|X)					Joint Percentages P(C|X) * P(HDF)					
	Reject	Accept	s	s2	s1	S	Check	s	s2	s1	S	Check	s+s2
s.iso_03	44%	56%	30%	23%	47%	0%	100%	17%	13%	26%	0%	56%	
s.iso_04	90%	10%	2%	82%	15%	1%	100%	0%	8%	2%	0%	10%	
s.iso_05	27%	73%	19%	1%	66%	14%	100%	14%	1%	48%	10%	73%	
s.iso_06	63%	37%	9%	1%	83%	7%	100%	3%	0%	31%	3%	37%	
s.iso_07	35%	65%	7%	2%	84%	7%	100%	5%	1%	55%	5%	65%	
s.iso_08	67%	33%	7%	2%	85%	6%	100%	2%	1%	28%	2%	33%	
s.iso_09	42%	58%	8%	2%	85%	5%	100%	5%	1%	49%	3%	58%	
s.iso_10	15%	85%	12%	1%	86%	1%	100%	10%	1%	73%	1%	85%	
s.iso_11	39%	61%	10%	1%	80%	9%	100%	6%	1%	49%	5%	61%	
Session 2,3	**47%**	**53%**	**12%**	**13%**	**70%**	**6%**	**100%**	**7%**	**3%**	**40%**	**3%**	**53%**	**10%**
s.iso_12	40%	60%	0%	98%	2%	0%	100%	0%	59%	1%	0%	60%	
s.iso_13	9%	91%	8%	1%	91%	0%	100%	7%	1%	83%	0%	91%	
s.iso_14	17%	83%	15%	1%	81%	3%	100%	12%	1%	67%	2%	83%	
s.iso_15	14%	86%	7%	3%	90%	0%	100%	6%	3%	77%	0%	86%	
s.iso_16	14%	86%	8%	1%	91%	0%	100%	7%	1%	78%	0%	86%	
s.iso_17	5%	95%	8%	3%	89%	0%	100%	8%	3%	85%	0%	95%	
s.iso_18	8%	92%	8%	4%	87%	1%	100%	7%	4%	80%	1%	92%	
Session 4	**15%**	**85%**	**8%**	**16%**	**76%**	**1%**	**100%**	**7%**	**10%**	**67%**	**0%**	**85%**	**17%**
s.iso_19	14%	86%	39%	4%	57%	0%	100%	34%	3%	49%	0%	86%	
s.iso_20	5%	95%	2%	2%	95%	1%	100%	2%	2%	90%	1%	95%	
s.iso_21	9%	91%	4%	3%	91%	2%	100%	4%	3%	83%	2%	91%	
s.iso_22	10%	90%	7%	30%	63%	0%	100%	6%	27%	57%	0%	90%	
s.iso_23	5%	95%	9%	13%	78%	0%	100%	9%	12%	74%	0%	95%	
s.iso_24	30%	70%	11%	5%	84%	0%	100%	8%	4%	59%	0%	70%	
Session 5	**12%**	**88%**	**12%**	**10%**	**78%**	**1%**	**100%**	**10%**	**8%**	**69%**	**0%**	**88%**	**19%**
s.iso_25	27%	73%	59%	9%	32%	0%	100%	43%	7%	23%	0%	73%	
s.iso_26	4%	96%	12%	31%	57%	0%	100%	12%	30%	55%	0%	96%	
s.iso_27	5%	95%	20%	6%	74%	0%	100%	19%	6%	70%	0%	95%	
s.iso_28	6%	94%	20%	46%	34%	0%	100%	19%	43%	32%	0%	94%	
Session 6	**11%**	**90%**	**28%**	**23%**	**49%**	**0%**	**100%**	**23%**	**21%**	**45%**	**0%**	**90%**	**44%**

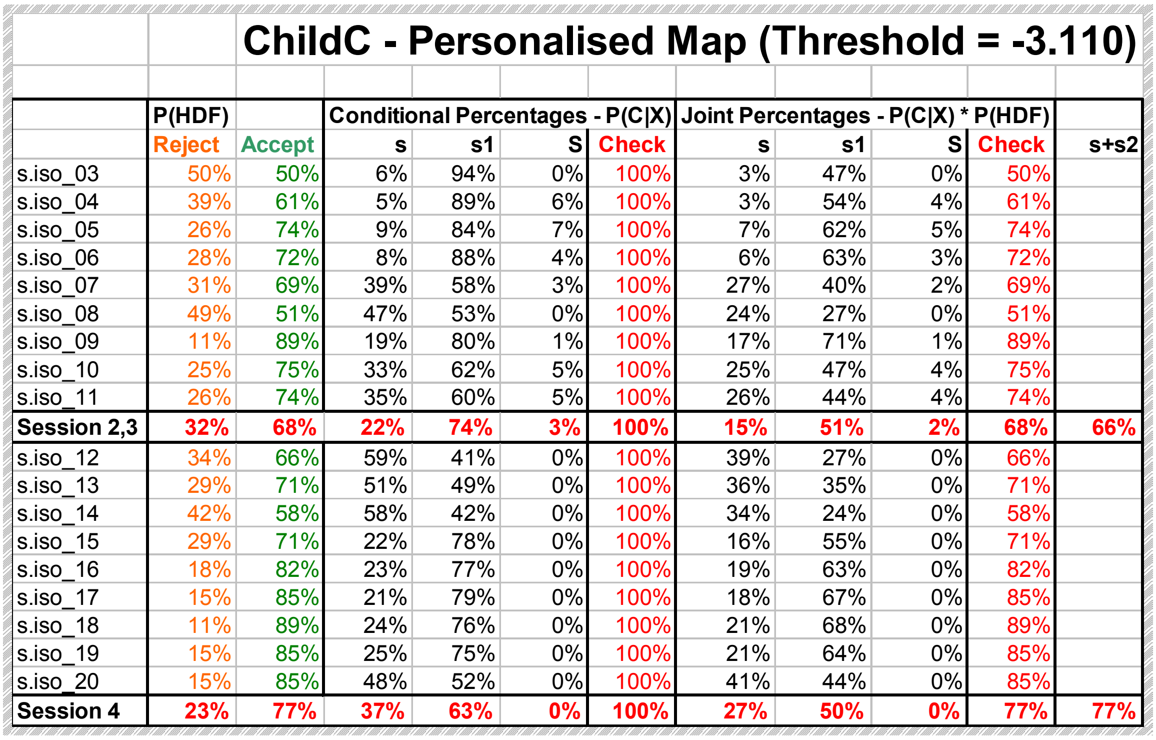
Table Appendices-8 *ChildA scoring per session on his individualised map.*



ChildB - Personalised Map (Threshold = -3.110)

	P(HDF)		Conditional Percentages - P(C|X)					Joint Percentages - P(C|X) * P(HDF)					
	Reject	Accept	s	s2	s1	S	Check	s	s2	s1	S	Check	s+s2
s.iso_03	5%	95%	0%	77%	23%	0%	100%	0%	73%	22%	0%	95%	
s.iso_04	17%	83%	0%	7%	93%	0%	100%	0%	6%	77%	0%	83%	
s.iso_05	17%	83%	3%	21%	76%	0%	100%	2%	17%	63%	0%	83%	
s.iso_06	8%	92%	1%	15%	84%	0%	100%	1%	14%	77%	0%	92%	
s.iso_07	27%	73%	3%	10%	81%	6%	100%	2%	7%	59%	4%	73%	
s.iso_08	10%	90%	2%	13%	85%	0%	100%	2%	12%	77%	0%	90%	
Session 2,3	**14%**	**86%**	**2%**	**24%**	**74%**	**1%**	**100%**	**1%**	**22%**	**63%**	**1%**	**86%**	**23%**
s.iso_10	18%	82%	2%	98%	0%	0%	100%	2%	80%	0%	0%	82%	
s.iso_11	17%	83%	2%	98%	0%	0%	100%	2%	81%	0%	0%	83%	
s.iso_12	12%	88%	1%	98%	1%	0%	100%	1%	86%	1%	0%	88%	
s.iso_13	18%	82%	3%	95%	2%	0%	100%	2%	78%	2%	0%	82%	
s.iso_14	8%	92%	5%	83%	12%	0%	100%	5%	76%	11%	0%	92%	
s.iso_15	23%	77%	3%	87%	10%	0%	100%	2%	67%	8%	0%	77%	
s.iso_16	11%	89%	3%	88%	9%	0%	100%	3%	78%	8%	0%	89%	
Session 4	**15%**	**85%**	**3%**	**92%**	**5%**	**0%**	**100%**	**2%**	**78%**	**4%**	**0%**	**85%**	**81%**
s.iso_17	16%	84%	12%	70%	18%	0%	100%	10%	59%	15%	0%	84%	
s.iso_18	18%	82%	17%	49%	34%	0%	100%	14%	40%	28%	0%	82%	
s.iso_19	24%	76%	7%	70%	23%	0%	100%	5%	53%	17%	0%	76%	
s.iso_20	20%	80%	12%	61%	27%	0%	100%	10%	49%	22%	0%	80%	
Session 5	**20%**	**81%**	**12%**	**63%**	**26%**	**0%**	**100%**	**10%**	**50%**	**21%**	**0%**	**81%**	**60%**
s.iso_21	7%	93%	11%	54%	35%	0%	100%	10%	50%	33%	0%	93%	
s.iso_22	10%	90%	12%	62%	26%	0%	100%	11%	56%	23%	0%	90%	
s.iso_23	7%	93%	16%	48%	36%	0%	100%	15%	45%	33%	0%	93%	
s.iso_24	31%	69%	5%	83%	12%	0%	100%	3%	57%	8%	0%	69%	
Session 6	**14%**	**86%**	**11%**	**62%**	**27%**	**0%**	**100%**	**10%**	**52%**	**24%**	**0%**	**86%**	**62%**

Table Appendices-9 *ChildB scoring per session on her individualised map.*

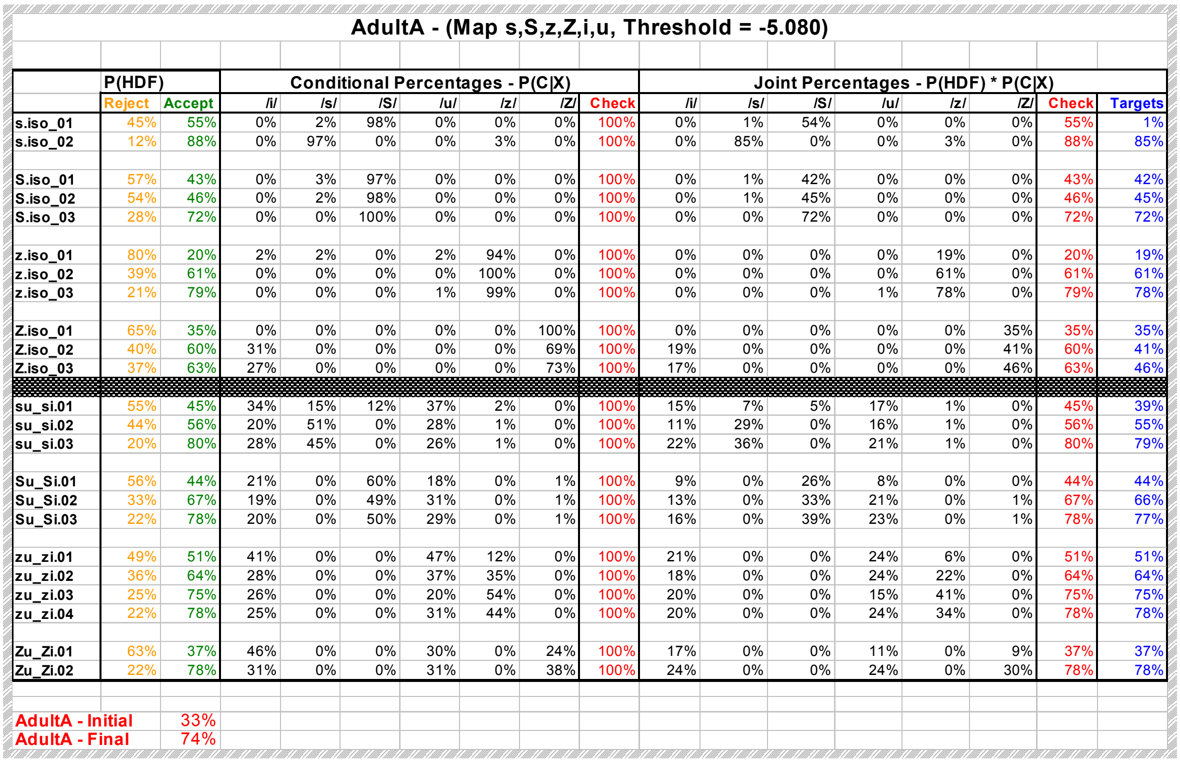


ChildC - Personalised Map (Threshold = -3.110)

	P(HDF)		Conditional Percentages - P(C|X)				Joint Percentages - P(C|X) * P(HDF)				
	Reject	Accept	s	s1	S	Check	s	s1	S	Check	s+s2
s.iso_03	50%	50%	6%	94%	0%	100%	3%	47%	0%	50%	
s.iso_04	39%	61%	5%	89%	6%	100%	3%	54%	4%	61%	
s.iso_05	26%	74%	9%	84%	7%	100%	7%	62%	5%	74%	
s.iso_06	28%	72%	8%	88%	4%	100%	6%	63%	3%	72%	
s.iso_07	31%	69%	39%	58%	3%	100%	27%	40%	2%	69%	
s.iso_08	49%	51%	47%	53%	0%	100%	24%	27%	0%	51%	
s.iso_09	11%	89%	19%	80%	1%	100%	17%	71%	1%	89%	
s.iso_10	25%	75%	33%	62%	5%	100%	25%	47%	4%	75%	
s.iso_11	26%	74%	35%	60%	5%	100%	26%	44%	4%	74%	
Session 2,3	**32%**	**68%**	**22%**	**74%**	**3%**	**100%**	**15%**	**51%**	**2%**	**68%**	**66%**
s.iso_12	34%	66%	59%	41%	0%	100%	39%	27%	0%	66%	
s.iso_13	29%	71%	51%	49%	0%	100%	36%	35%	0%	71%	
s.iso_14	42%	58%	58%	42%	0%	100%	34%	24%	0%	58%	
s.iso_15	29%	71%	22%	78%	0%	100%	16%	55%	0%	71%	
s.iso_16	18%	82%	23%	77%	0%	100%	19%	63%	0%	82%	
s.iso_17	15%	85%	21%	79%	0%	100%	18%	67%	0%	85%	
s.iso_18	11%	89%	24%	76%	0%	100%	21%	68%	0%	89%	
s.iso_19	15%	85%	25%	75%	0%	100%	21%	64%	0%	85%	
s.iso_20	15%	85%	48%	52%	0%	100%	41%	44%	0%	85%	
Session 4	**23%**	**77%**	**37%**	**63%**	**0%**	**100%**	**27%**	**50%**	**0%**	**77%**	**77%**

Table Appendices-10 *ChildC scoring per session on his individualised map.*

VI Accent Modification Results



AdultA - (Map s,S,z,Z,i,u, Threshold = -5.080)

	P(HDF)		Conditional Percentages - P(C|X)							Joint Percentages - P(HDF) * P(C|X)							
	Reject	Accept	/i/	/s/	/S/	/u/	/z/	/Z/	Check	/i/	/s/	/S/	/u/	/z/	/Z/	Check	Targets
s.iso_01	45%	55%	0%	2%	98%	0%	0%	0%	100%	0%	1%	54%	0%	0%	0%	55%	1%
s.iso_02	12%	88%	0%	97%	0%	0%	3%	0%	100%	0%	85%	0%	0%	3%	0%	88%	85%
S.iso_01	57%	43%	0%	3%	97%	0%	0%	0%	100%	0%	1%	42%	0%	0%	0%	43%	42%
S.iso_02	54%	46%	0%	2%	98%	0%	0%	0%	100%	0%	1%	45%	0%	0%	0%	46%	45%
S.iso_03	28%	72%	0%	0%	100%	0%	0%	0%	100%	0%	0%	72%	0%	0%	0%	72%	72%
z.iso_01	80%	20%	2%	2%	0%	2%	94%	0%	100%	0%	0%	0%	0%	19%	0%	20%	19%
z.iso_02	39%	61%	0%	0%	0%	0%	100%	0%	100%	0%	0%	0%	0%	61%	0%	61%	61%
z.iso_03	21%	79%	0%	0%	0%	1%	99%	0%	100%	0%	0%	0%	1%	78%	0%	79%	78%
Z.iso_01	65%	35%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	0%	35%	35%	35%
Z.iso_02	40%	60%	31%	0%	0%	0%	0%	69%	100%	19%	0%	0%	0%	0%	41%	60%	41%
Z.iso_03	37%	63%	27%	0%	0%	0%	0%	73%	100%	17%	0%	0%	0%	0%	46%	63%	46%
su_si.01	55%	45%	34%	15%	12%	37%	2%	0%	100%	15%	7%	5%	17%	1%	0%	45%	39%
su_si.02	44%	56%	20%	51%	0%	28%	1%	0%	100%	11%	29%	0%	16%	1%	0%	56%	55%
su_si.03	20%	80%	28%	45%	0%	26%	1%	0%	100%	22%	36%	0%	21%	1%	0%	80%	79%
Su_Si.01	56%	44%	21%	0%	60%	18%	0%	1%	100%	9%	0%	26%	8%	0%	0%	44%	44%
Su_Si.02	33%	67%	19%	0%	49%	31%	0%	1%	100%	13%	0%	33%	21%	0%	1%	67%	66%
Su_Si.03	22%	78%	20%	0%	50%	29%	0%	1%	100%	16%	0%	39%	23%	0%	1%	78%	77%
zu_zi.01	49%	51%	41%	0%	0%	47%	12%	0%	100%	21%	0%	0%	24%	6%	0%	51%	51%
zu_zi.02	36%	64%	28%	0%	0%	37%	35%	0%	100%	18%	0%	0%	24%	22%	0%	64%	64%
zu_zi.03	25%	75%	26%	0%	0%	20%	54%	0%	100%	20%	0%	0%	15%	41%	0%	75%	75%
zu_zi.04	22%	78%	25%	0%	0%	31%	44%	0%	100%	20%	0%	0%	24%	34%	0%	78%	78%
Zu_Zi.01	63%	37%	46%	0%	0%	30%	0%	24%	100%	17%	0%	0%	11%	0%	9%	37%	37%
Zu_Zi.02	22%	78%	31%	0%	0%	31%	0%	38%	100%	24%	0%	0%	24%	0%	30%	78%	78%

AdultA - Initial	33%
AdultA - Final	74%

Table Appendices-11 *Accent performance of AdultA on 'iso4vo2frlvq' map*

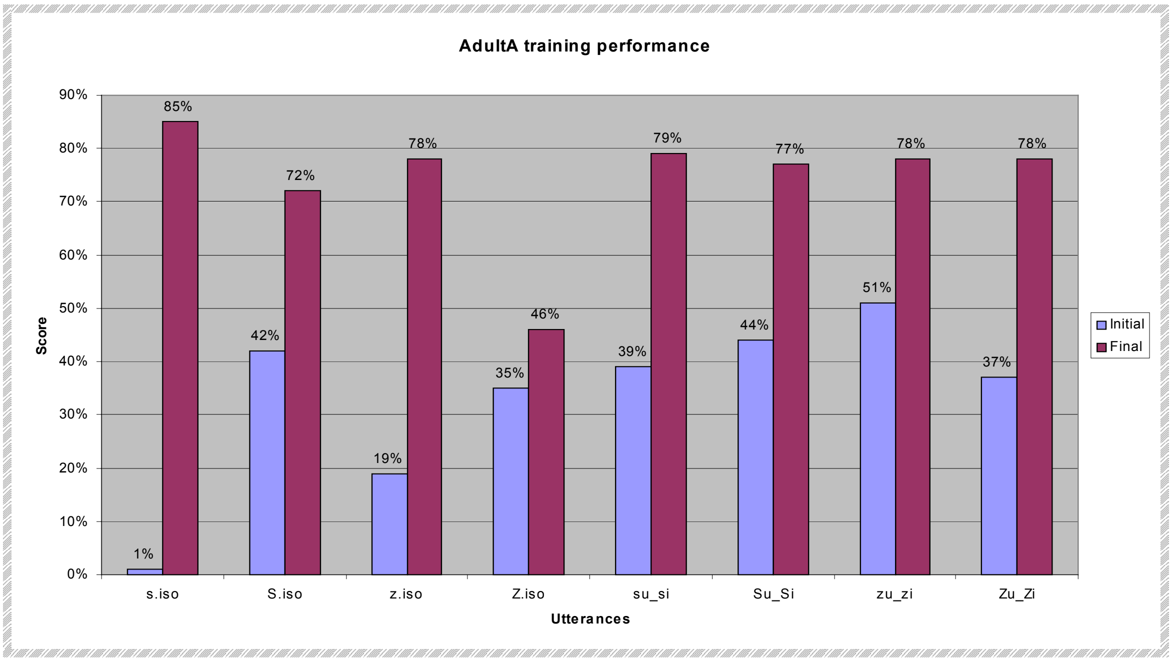
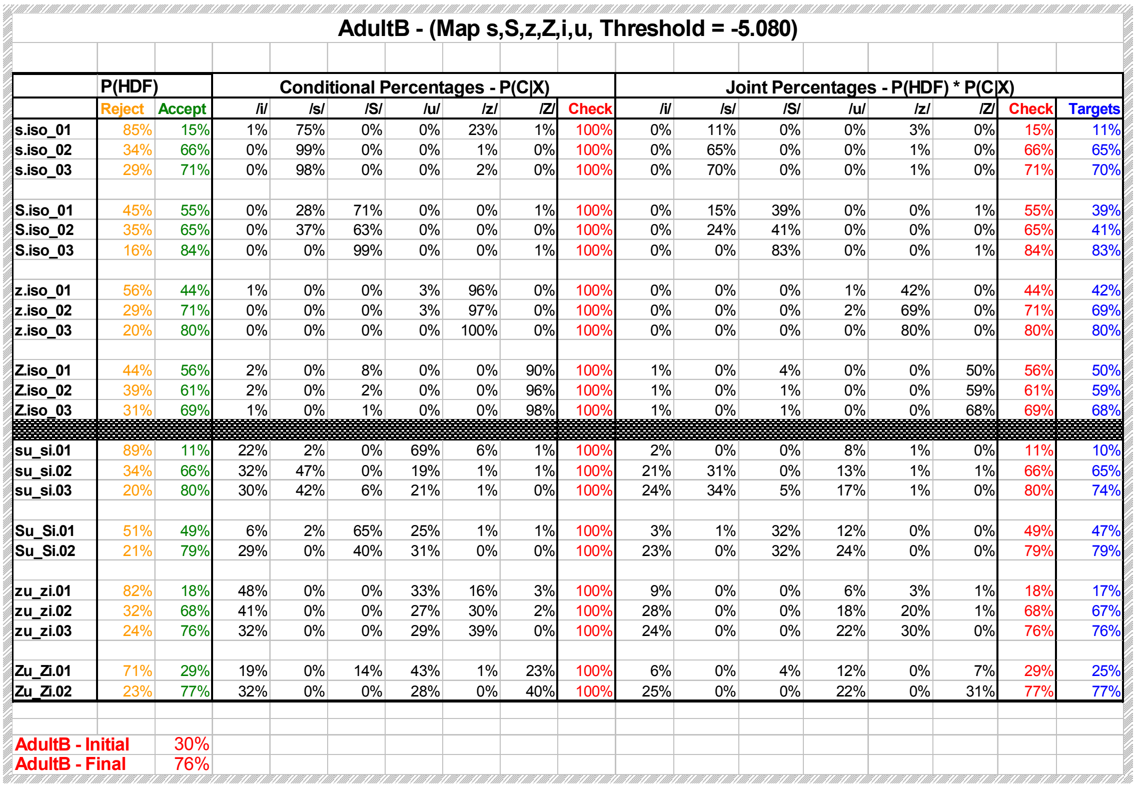


Figure Appendices-36 *AdultA accent comparison of initial and final session*



AdultB - (Map s,S,z,Z,i,u, Threshold = -5.080)																	
	P(HDF)		Conditional Percentages - P(C|X)							Joint Percentages - P(HDF) * P(C|X)							
	Reject	Accept	/i/	/s/	/S/	/u/	/z/	/Z/	Check	/i/	/s/	/S/	/u/	/z/	/Z/	Check	Targets
s.iso_01	85%	15%	1%	75%	0%	0%	23%	1%	100%	0%	11%	0%	0%	3%	0%	15%	11%
s.iso_02	34%	66%	0%	99%	0%	0%	1%	0%	100%	0%	65%	0%	0%	1%	0%	66%	65%
s.iso_03	29%	71%	0%	98%	0%	0%	2%	0%	100%	0%	70%	0%	0%	1%	0%	71%	70%
S.iso_01	45%	55%	0%	28%	71%	0%	0%	1%	100%	0%	15%	39%	0%	0%	1%	55%	39%
S.iso_02	35%	65%	0%	37%	63%	0%	0%	0%	100%	0%	24%	41%	0%	0%	0%	65%	41%
S.iso_03	16%	84%	0%	0%	99%	0%	0%	1%	100%	0%	0%	83%	0%	0%	1%	84%	83%
z.iso_01	56%	44%	1%	0%	0%	3%	96%	0%	100%	0%	0%	0%	1%	42%	0%	44%	42%
z.iso_02	29%	71%	0%	0%	0%	3%	97%	0%	100%	0%	0%	0%	2%	69%	0%	71%	69%
z.iso_03	20%	80%	0%	0%	0%	0%	100%	0%	100%	0%	0%	0%	0%	80%	0%	80%	80%
Z.iso_01	44%	56%	2%	0%	8%	0%	0%	90%	100%	1%	0%	4%	0%	0%	50%	56%	50%
Z.iso_02	39%	61%	2%	0%	2%	0%	0%	96%	100%	1%	0%	1%	0%	0%	59%	61%	59%
Z.iso_03	31%	69%	1%	0%	1%	0%	0%	98%	100%	1%	0%	1%	0%	0%	68%	69%	68%
su_si.01	89%	11%	22%	2%	0%	69%	6%	1%	100%	2%	0%	0%	8%	1%	0%	11%	10%
su_si.02	34%	66%	32%	47%	0%	19%	1%	1%	100%	21%	31%	0%	13%	1%	1%	66%	65%
su_si.03	20%	80%	30%	42%	6%	21%	1%	0%	100%	24%	34%	5%	17%	1%	0%	80%	74%
Su_Si.01	51%	49%	6%	2%	65%	25%	1%	1%	100%	3%	1%	32%	12%	0%	0%	49%	47%
Su_Si.02	21%	79%	29%	0%	40%	31%	0%	0%	100%	23%	0%	32%	24%	0%	0%	79%	79%
zu_zi.01	82%	18%	48%	0%	0%	33%	16%	3%	100%	9%	0%	0%	6%	3%	1%	18%	17%
zu_zi.02	32%	68%	41%	0%	0%	27%	30%	2%	100%	28%	0%	0%	18%	20%	1%	68%	67%
zu_zi.03	24%	76%	32%	0%	0%	29%	39%	0%	100%	24%	0%	0%	22%	30%	0%	76%	76%
Zu_Zi.01	71%	29%	19%	0%	14%	43%	1%	23%	100%	6%	0%	4%	12%	0%	7%	29%	25%
Zu_Zi.02	23%	77%	32%	0%	0%	28%	0%	40%	100%	25%	0%	0%	22%	0%	31%	77%	77%
AdultB - Initial		30%															
AdultB - Final		76%															

Table Appendices-12 *Accent performance of AdultB on 'iso4vo2frlvq' map*

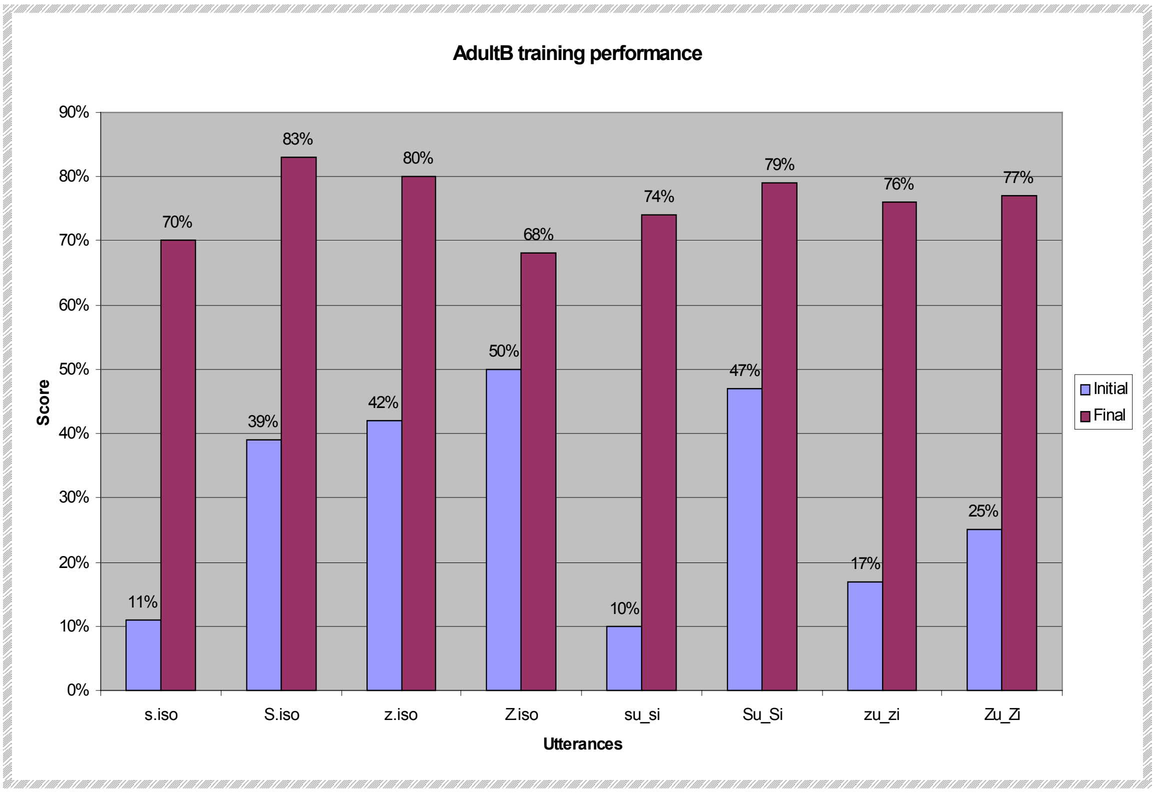
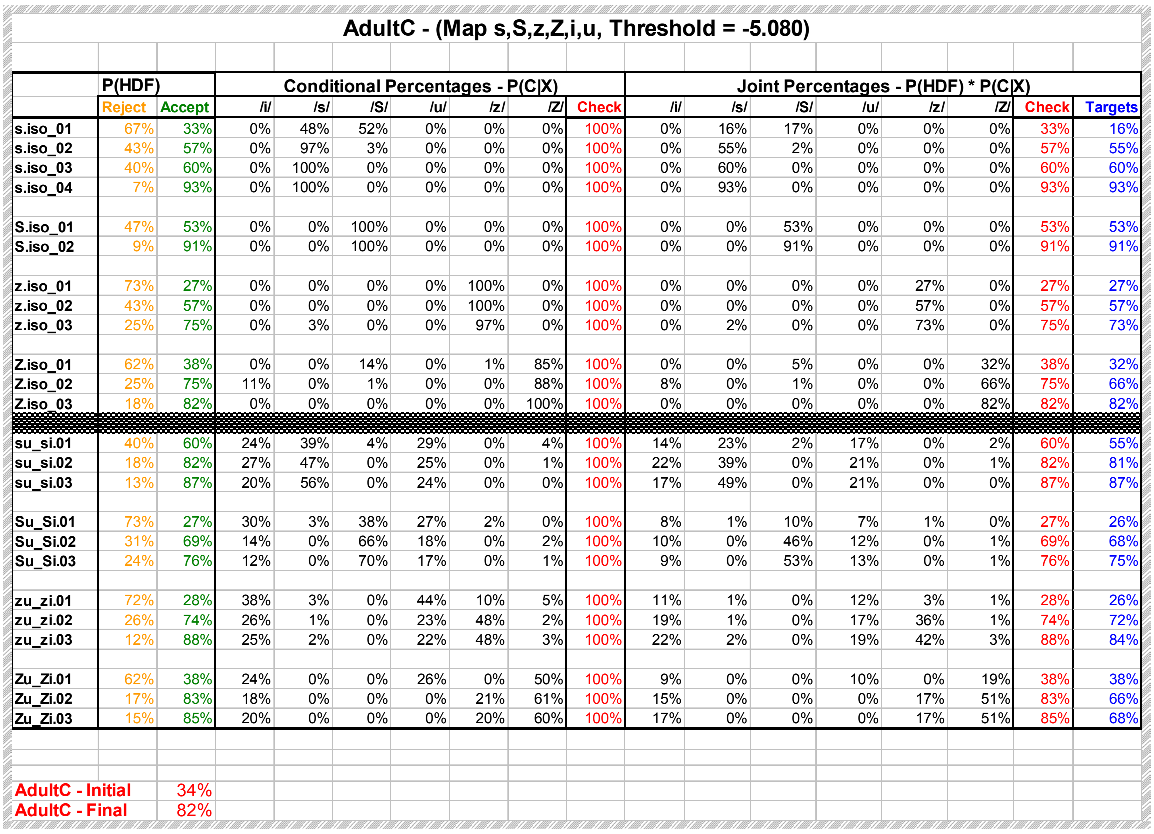


Figure Appendices-37 *AdultB accent comparison of initial and final session*



AdultC - (Map s,S,z,Z,i,u, Threshold = -5.080)

	P(HDF)		Conditional Percentages - P(C|X)							Joint Percentages - P(HDF) * P(C|X)							
	Reject	Accept	/i/	/s/	/S/	/u/	/z/	/Z/	Check	/i/	/s/	/S/	/u/	/z/	/Z/	Check	Targets
s.iso_01	67%	33%	0%	48%	52%	0%	0%	0%	100%	0%	16%	17%	0%	0%	0%	33%	16%
s.iso_02	43%	57%	0%	97%	3%	0%	0%	0%	100%	0%	55%	2%	0%	0%	0%	57%	55%
s.iso_03	40%	60%	0%	100%	0%	0%	0%	0%	100%	0%	60%	0%	0%	0%	0%	60%	60%
s.iso_04	7%	93%	0%	100%	0%	0%	0%	0%	100%	0%	93%	0%	0%	0%	0%	93%	93%
S.iso_01	47%	53%	0%	0%	100%	0%	0%	0%	100%	0%	0%	53%	0%	0%	0%	53%	53%
S.iso_02	9%	91%	0%	0%	100%	0%	0%	0%	100%	0%	0%	91%	0%	0%	0%	91%	91%
z.iso_01	73%	27%	0%	0%	0%	0%	100%	0%	100%	0%	0%	0%	0%	27%	0%	27%	27%
z.iso_02	43%	57%	0%	0%	0%	0%	100%	0%	100%	0%	0%	0%	0%	57%	0%	57%	57%
z.iso_03	25%	75%	0%	3%	0%	0%	97%	0%	100%	0%	2%	0%	0%	73%	0%	75%	73%
Z.iso_01	62%	38%	0%	0%	14%	0%	1%	85%	100%	0%	0%	5%	0%	0%	32%	38%	32%
Z.iso_02	25%	75%	11%	0%	1%	0%	0%	88%	100%	8%	0%	1%	0%	0%	66%	75%	66%
Z.iso_03	18%	82%	0%	0%	0%	0%	0%	100%	100%	0%	0%	0%	0%	0%	82%	82%	82%
su_si.01	40%	60%	24%	39%	4%	29%	0%	4%	100%	14%	23%	2%	17%	0%	2%	60%	55%
su_si.02	18%	82%	27%	47%	0%	25%	0%	1%	100%	22%	39%	0%	21%	0%	1%	82%	81%
su_si.03	13%	87%	20%	56%	0%	24%	0%	0%	100%	17%	49%	0%	21%	0%	0%	87%	87%
Su_Si.01	73%	27%	30%	3%	38%	27%	2%	0%	100%	8%	1%	10%	7%	1%	0%	27%	26%
Su_Si.02	31%	69%	14%	0%	66%	18%	0%	2%	100%	10%	0%	46%	12%	0%	1%	69%	68%
Su_Si.03	24%	76%	12%	0%	70%	17%	0%	1%	100%	9%	0%	53%	13%	0%	1%	76%	75%
zu_zi.01	72%	28%	38%	3%	0%	44%	10%	5%	100%	11%	1%	0%	12%	3%	1%	28%	26%
zu_zi.02	26%	74%	26%	1%	0%	23%	48%	2%	100%	19%	1%	0%	17%	36%	1%	74%	72%
zu_zi.03	12%	88%	25%	2%	0%	22%	48%	3%	100%	22%	2%	0%	19%	42%	3%	88%	84%
Zu_Zi.01	62%	38%	24%	0%	0%	26%	0%	50%	100%	9%	0%	0%	10%	0%	19%	38%	38%
Zu_Zi.02	17%	83%	18%	0%	0%	0%	21%	61%	100%	15%	0%	0%	0%	17%	51%	83%	66%
Zu_Zi.03	15%	85%	20%	0%	0%	0%	20%	60%	100%	17%	0%	0%	0%	17%	51%	85%	68%
AdultC - Initial		34%															
AdultC - Final		82%															

Table Appendices-13 *Accent performance of AdultC on 'iso4vo2frlvq' map*

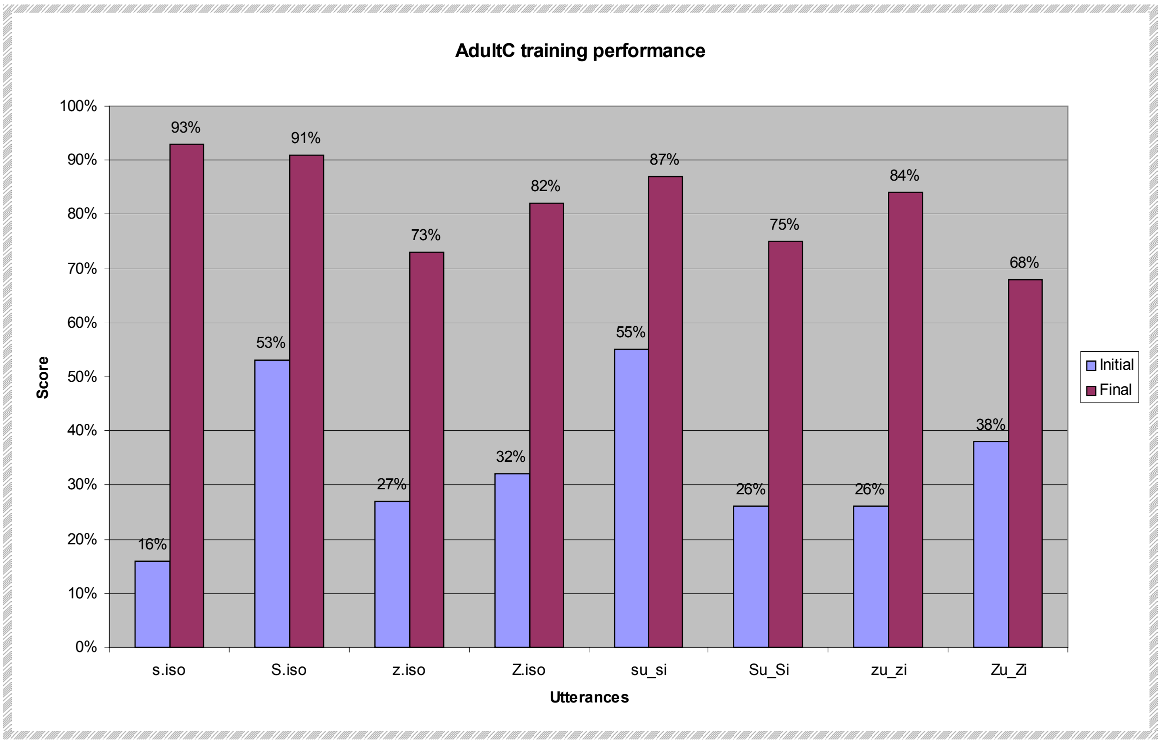


Figure Appendices-38 *AdultC accent comparison of initial and final session*

AdultD - (Map s,S,z,Z,i,u, Threshold = -5.080)

	P(HDF)		Conditional Percentages - P(C|X)							Joint Percentages - P(HDF) * P(C|X)							
	Reject	Accept	/i/	/s/	/S/	/u/	/z/	/Z/	Check	/i/	/s/	/S/	/u/	/z/	/Z/	Check	Targets
s.iso_01	98%	2%	0%	30%	67%	0%	0%	3%	100%	0%	1%	1%	0%	0%	0%	2%	1%
s.iso_02	95%	5%	0%	100%	0%	0%	0%	0%	100%	0%	5%	0%	0%	0%	0%	5%	5%
s.iso_03	16%	84%	0%	99%	1%	0%	0%	0%	100%	0%	83%	1%	0%	0%	0%	84%	83%
S.iso_01	60%	40%	0%	0%	99%	0%	0%	1%	100%	0%	0%	40%	0%	0%	0%	40%	40%
S.iso_02	47%	53%	0%	0%	100%	0%	0%	0%	100%	0%	0%	53%	0%	0%	0%	53%	53%
S.iso_03	41%	59%	0%	1%	99%	0%	0%	0%	100%	0%	1%	58%	0%	0%	0%	59%	58%
S.iso_04	38%	62%	0%	0%	100%	0%	0%	0%	100%	0%	0%	62%	0%	0%	0%	62%	62%
z.iso_01	98%	2%	23%	0%	0%	0%	54%	23%	100%	0%	0%	0%	0%	1%	0%	2%	1%
z.iso_02	74%	26%	1%	0%	0%	4%	95%	0%	100%	0%	0%	0%	1%	25%	0%	26%	25%
Z.iso_01	75%	25%	1%	0%	0%	0%	0%	99%	100%	0%	0%	0%	0%	0%	25%	25%	25%
Z.iso_02	48%	52%	39%	0%	0%	1%	0%	60%	100%	20%	0%	0%	1%	0%	31%	52%	31%
su_si.01	64%	36%	7%	2%	36%	26%	0%	29%	100%	3%	1%	13%	9%	0%	10%	36%	13%
su_si.02	15%	85%	26%	47%	0%	27%	0%	0%	100%	22%	40%	0%	23%	0%	0%	85%	85%
Su_Si.01	53%	47%	13%	0%	35%	31%	0%	21%	100%	6%	0%	16%	15%	0%	10%	47%	37%
Su_Si.02	25%	75%	15%	0%	54%	22%	0%	9%	100%	11%	0%	41%	17%	0%	7%	75%	68%
zu_zi.01	70%	30%	23%	0%	0%	67%	8%	2%	100%	7%	0%	0%	20%	2%	1%	30%	29%
zu_zi.02	49%	51%	28%	0%	0%	53%	0%	19%	100%	14%	0%	0%	27%	0%	10%	51%	41%
zu_zi.03	29%	71%	29%	0%	0%	30%	41%	0%	100%	21%	0%	0%	21%	29%	0%	71%	71%
Zu_Zi.01	62%	38%	24%	0%	0%	21%	0%	54%	99%	9%	0%	0%	8%	0%	21%	38%	38%
Zu_Zi.02	57%	43%	30%	0%	0%	39%	0%	31%	100%	13%	0%	0%	17%	0%	13%	43%	43%
AdultD - Initial		23%															
AdultD - Final		59%															

Table Appendices-14 *Accent performance of AdultD on 'iso4vo2frlvq' map*

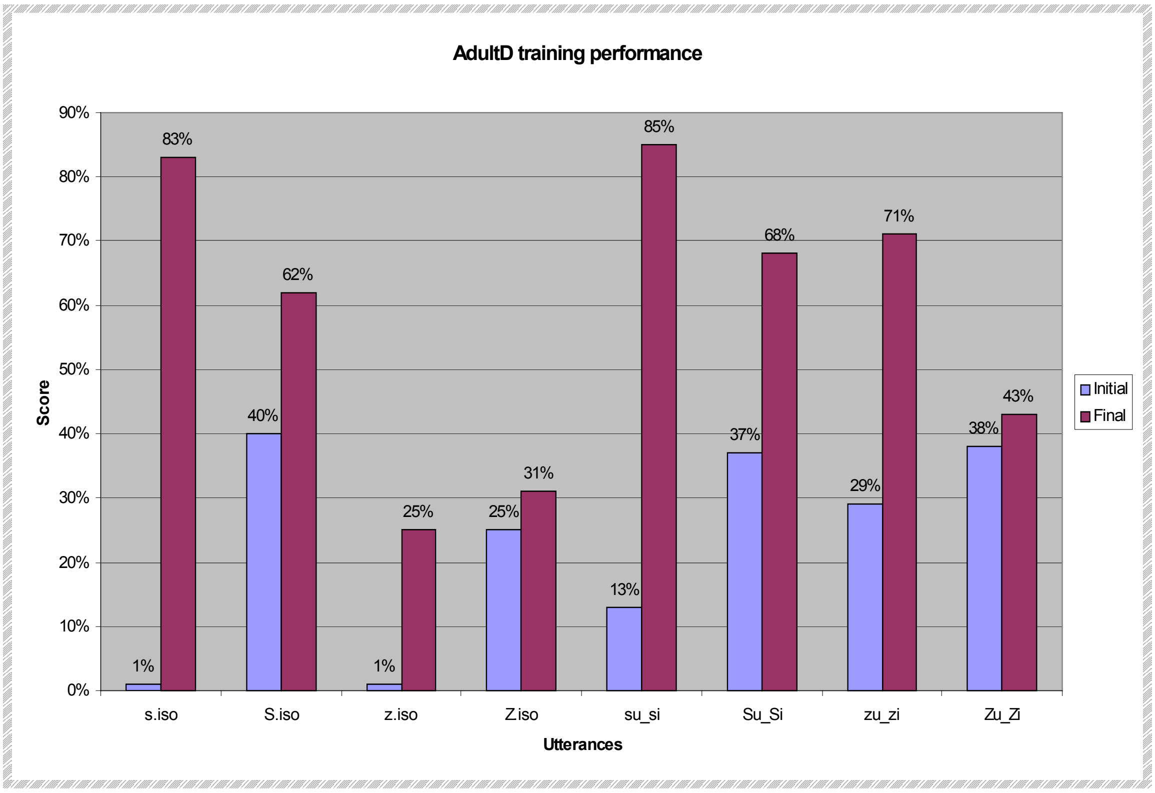


Figure Appendices-39 *AdultD accent comparison of initial and final session*

VII Maps Configuration

File name	*'ch-iso3vo3frlvq'*					
Description	Normal children map with three vowels and three sibilant Fricative targets and an LVQ codebook set.					
Subjects analysed	18 normal children – 9 male and 9 female					
Data analysed	Vowels and consonants in isolation					
Classes	**Train set**	**Test set**	**LVQ set**			
/i/	2400	600	100			
/o/	2400	600	100			
/u/	2400	557	100			
/s/	2400	576	100			
/ʃ/	2400	600	100			
/z/	2400	600	100			
Total	**14400**	**3533**	**600**			
Net type	**Train set**	**Test set**	**Accuracy**	**Inp**	**Out**	**Hid**
SMANN	0.0012	0.0013	MSSE	9	2	16
PPANN	0.0209	0.0212	Entropy	9	6	16
	99 %	99 %	Classify			
KNN-1 ON LVQ	99 %	99 %	Classify			

Table Appendices-15 *Map configuration for 'ch-iso3vo3frlvq' file*

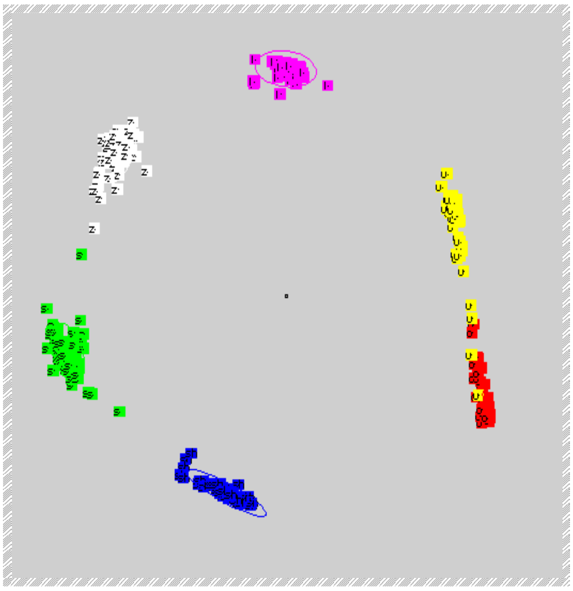


Figure Appendices-41 *Mapping of 'ch-iso3vo3frlvq' LVQ codebook set.*

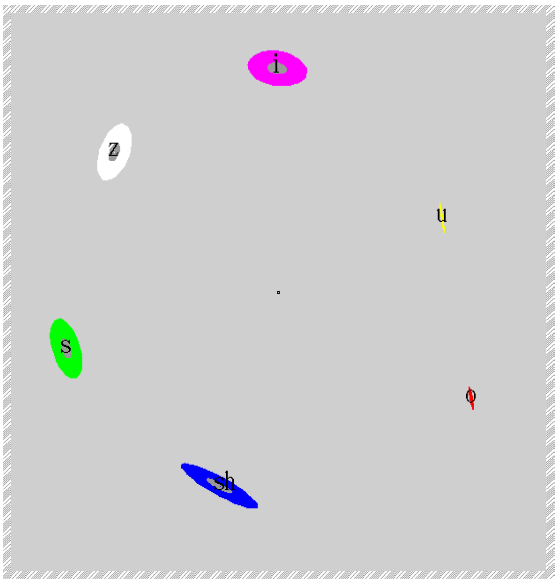


Figure Appendices-42 *Map representation of 'ch-iso3vo3frlvq' with ellipses.*

File name	*'ch-s-sh-i-u-o-lvq'*					
Description	Normal children map with three vowels and two sibilant fricative targets and an LVQ codebook set.					
Subjects analysed	18 normal children – 9 male and 9 female					
Data analysed	Vowels and consonants in isolation					
Classes	**Train set**	**Test set**	**LVQ set**			
/i/	2400	600	100			
/o/	2400	600	100			
/u/	2400	557	100			
/s/	2400	576	100			
/ʃ/	2400	600	100			
Total	**12000**	**2933**	**500**			
Net type	**Train set**	**Test set**	**Accuracy**	**Inp**	**Out**	**Hid**
SMANN	0.0016	0.0016	MSSE	9	2	32
PPANN	0.0198	0.0246	Entropy	9	5	16
	99 %	99 %	Classify			
KNN-1 ON LVQ	99 %	97 %	Classify			

Table Appendices-16 *Map configuration for 'ch-s-sh-i-u-o-lvq' file*

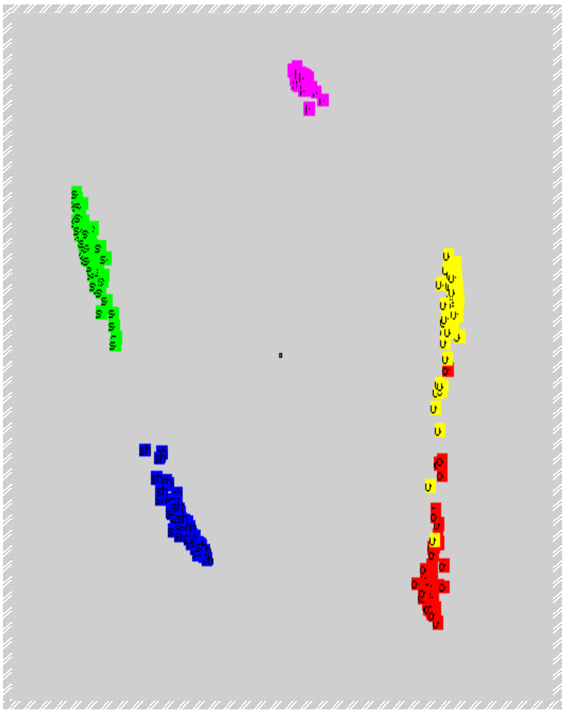


Figure Appendices-43 *Mapping of 'ch-s-sh-i-o-u-lvq' LVQ codebook set.*

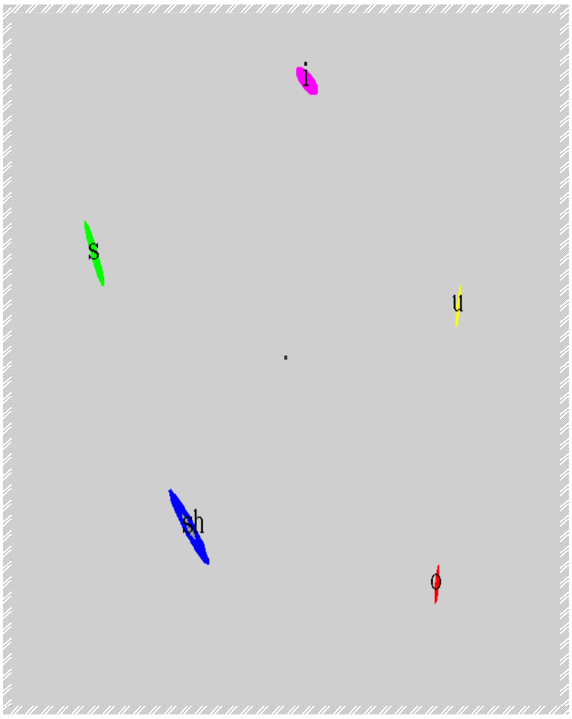


Figure Appendices-44 *Map representation of 'ch-s-sh-i-o-u-lvq' with ellipses.*

File name	'ChildB_s1s2_lvq'					
Description	Individualised map including two extra targets for improved skills in articulation of /s/ and an LVQ codebook set.					
Subjects analysed	18 normal children – 9 male and 9 female and the client					
Data analysed	Vowels and consonants in isolation and utterances of the client that appear to have consistently improved configurations for the /s/ sound.					
Classes	**Train set**	**Test set**	**LVQ set**			
/i/	2000	500	100			
/o/	2000	500	100			
/s/	2000	500	100			
/s1/	2000	500	100			
/s2/	2000	500	100			
/ʃ/	2000	500	100			
Total	**12000**	**3000**	**600**			
Net type	**Train set**	**Test set**	**Accuracy**	**Inp**	**Out**	**Hid**
SMANN	0.0036	0.0038	MSSE	9	2	16
PPANN	0.1163	0.1597	Entropy	9	6	64
	96 %	94 %	Classify			
KNN-1 ON LVQ	94 %	91 %	Classify			

Table Appendices-17 *Map configuration for 'ChildB_s1s2_lvq' file*

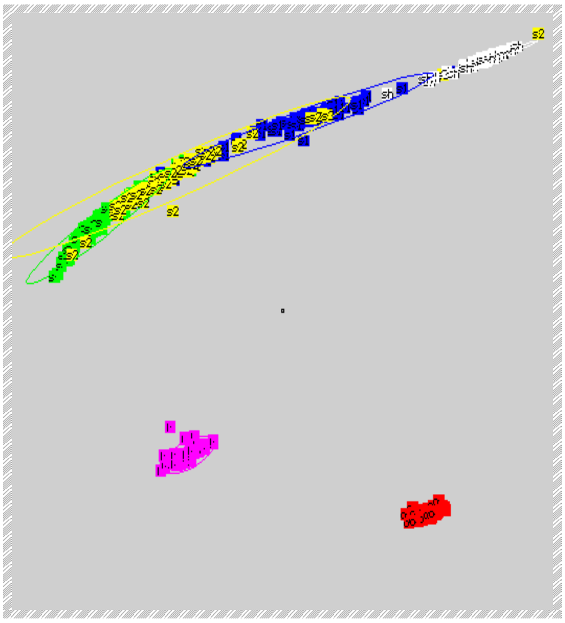


Figure Appendices-44 *Mapping of 'ChildB_s1s2_lvq' LVQ codebook set.*

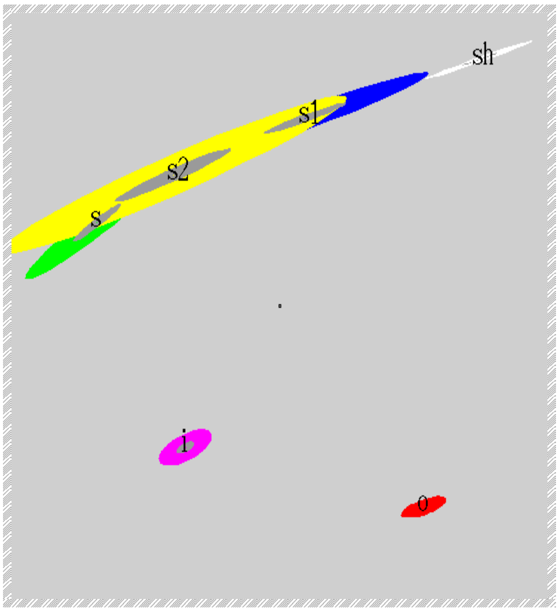


Figure Appendices-45 *Map representation of 'ChildB_s1s2_lvq' with ellipses.*

File name	***'ChildA_s1s2_lvq'***					
Description	Individualised map including two extra targets for improved skills in articulation of /s/ and an LVQ codebook set.					
Subjects analysed	18 normal children – 9 male and 9 female and the client					
Data analysed	Vowels and consonants in isolation and utterances of the client that appear to have consistently improved configurations for the /s/ sound.					
Classes	**Train set**	**Test set**	**LVQ set**			
/i/	2000	500	100			
/o/	2000	500	100			
/s/	2000	500	100			
/s1/	2000	500	100			
/s2/	2000	500	100			
/ʃ/	2000	500	100			
Total	**12000**	**3000**	**600**			
Net type	**Train set**	**Test set**	**Accuracy**	**Inp**	**Out**	**Hid**
SMANN	0.0036	0.0039	MSSE	9	2	16
PPANN	0.0990	0.1558	Entropy	9	6	32
	96 %	95 %	Classify			
KNN-1 ON LVQ	96 %	93 %	Classify			

Table Appendices-18 *Map configuration for 'ChildA_s1s2_lvq' file*

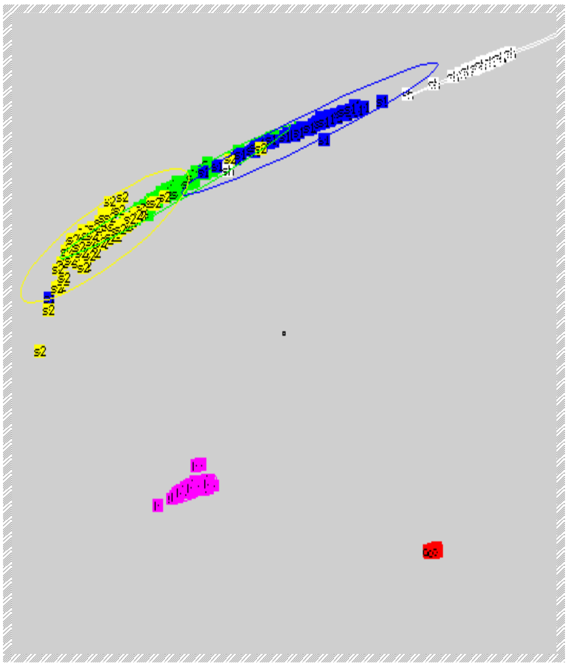


Figure Appendices-46 *Mapping of 'ChildA_s1s2_lvq' LVQ codebook set.*

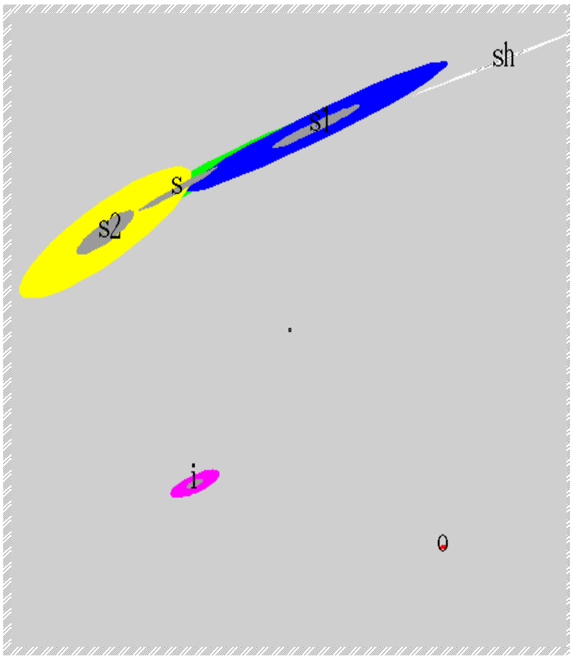


Figure Appendices-47 *Map representation of 'ChildA_s1s2_lvq' with ellipses.*

File name	***‘ChildC_s1_lvq’***					
Description	Individualised map including an extra target for improved skills in articulation of /s/ and an LVQ codebook set.					
Subjects analysed	18 normal children – 9 male and 9 female and the client					
Data analysed	Vowels and consonants in isolation and utterances of the client that appear to have a consistently improved configuration for the /s/ sound.					
Classes	**Train set**	**Test set**	**LVQ set**			
/i/	2000	500	100			
/o/	2000	500	100			
/s/	2000	500	100			
/s1/	2000	500	100			
/ʃ/	2000	500	100			
Total	**10000**	**2500**	**500**			
Net type	**Train set**	**Test set**	**Accuracy**	**Inp**	**Out**	**Hid**
SMANN	0.0047	0.0060	MSSE	9	2	16
PPANN	0.0531	0.0782	Entropy	9	5	32
	98 %	97 %	Classify			
KNN-1 ON LVQ	97 %	96 %	Classify			

Table Appendices-19 *Map configuration for ‘ChildC_s1_lvq’ file*

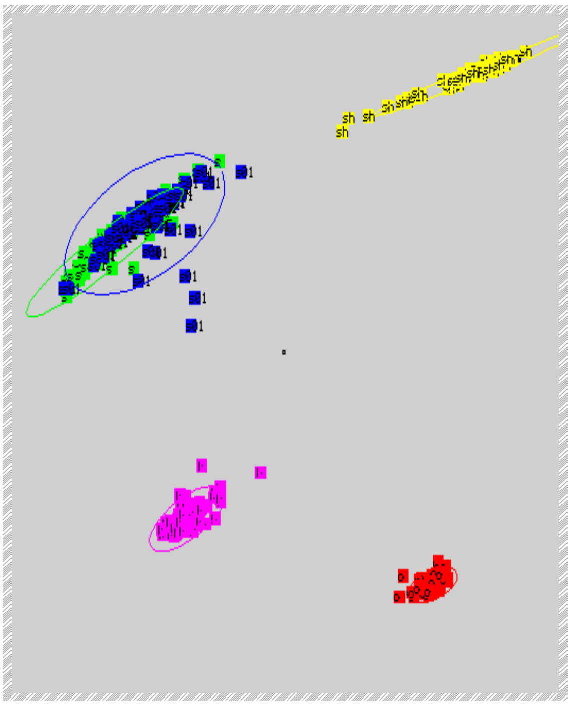


Figure Appendices-48 *Mapping of ‘ChildC_s1_lvq’ LVQ codebook set.*

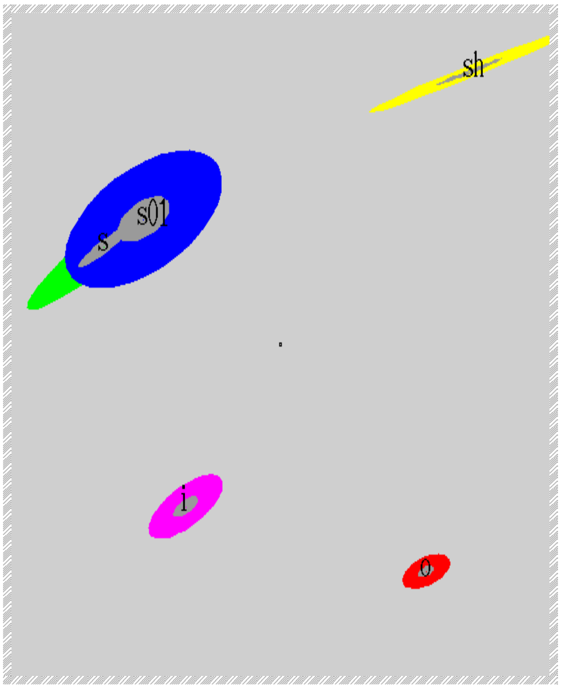


Figure Appendices-49 *Map representation of ‘ChildC_s1_lvq’ with ellipses.*

File name	'iso2vo4frlvq'					
Description	Normal English male adults map with two vowels and four sibilant fricative targets and an LVQ codebook set.					
Subjects analysed	8 normal English male adults					
Data analysed	Vowels and consonants in isolation					
Classes	**Train set**	**Test set**	**LVQ set**			
/i/	1200	300	100			
/u/	1200	300	100			
/s/	1200	300	100			
/ʃ/	1200	300	100			
/z/	1200	300	100			
/ʒ/	1200	300	100			
Total	**7200**	**1800**	**600**			
Net type	**Train set**	**Test set**	**Accuracy**	**Inp**	**Out**	**Hid**
SMANN	0.0006	0.0006	MSSE	9	2	16
PPANN	0.0013	0.0017	Entropy	9	6	16
	99.99 %	99.94 %	Classify			
KNN-1 ON LVQ	99.90 %	99.89 %	Classify			

Table Appendices-20 *Map configuration for 'iso2vo4frlvq' file*

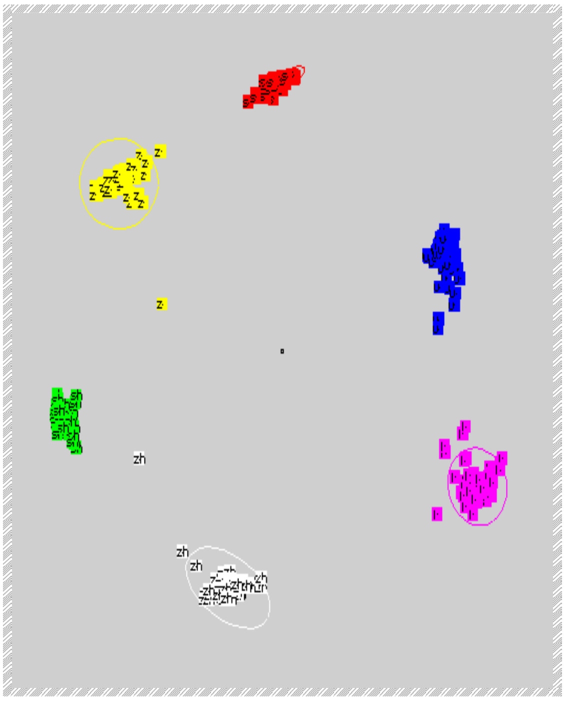


Figure Appendices-50 *Mapping of 'iso2vo4frlvq' LVQ codebook set.*

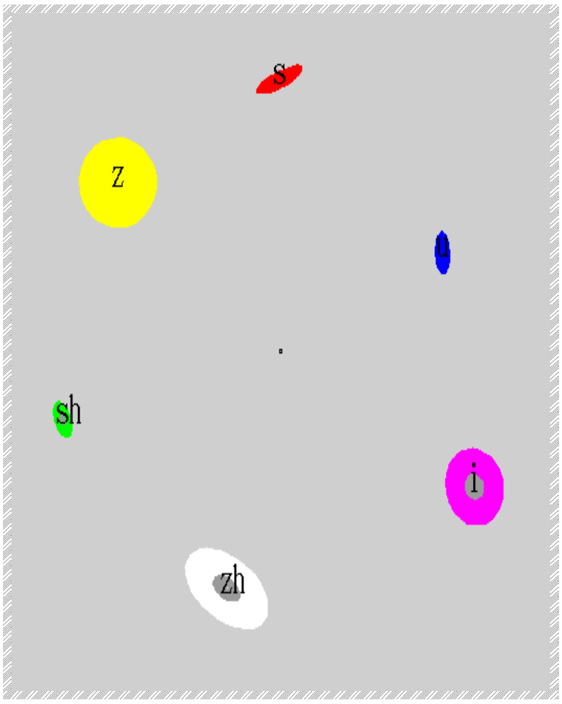


Figure Appendices-51 *Map representation of 'iso2vo4frlvq' phone classes as ellipses*

File name	***'freeXu_lvq'***					
Description	Normal English male adults map with two vowels and four sibilant fricative targets and an LVQ codebook set.					
Subjects analysed	6 normal English male adults – 8 repetitions each					
Data analysed	Vowels and consonants in i X u context, X is (s, ʃ, z, ʒ).					
Classes	**Train set**	**Test set**	**LVQ set**			
/i/	800	200	100			
/u/	800	200	100			
/s/	800	200	100			
/ʃ/	800	200	100			
/z/	800	127	100			
/ʒ/	800	150	100			
Total	**4800**	**1077**	**600**			
Net type	**Train set**	**Test set**	**Accuracy**	**Inp**	**Out**	**Hid**
SMANN	0.0117	0.0121	MSSE	9	2	64
PPANN	0.1799	0.2757	Entropy	9	6	64
	93.06%	89.60 %	Classify			
KNN-1 ON LVQ	92.27%	88.58 %	Classify			

Table Appendices-21 *Map configuration for 'freeXu_lvq' file*

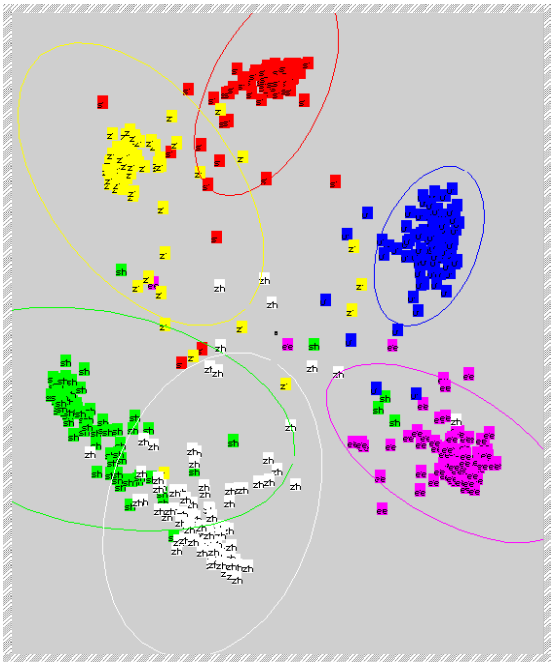


Figure Appendices-52 *Mapping of 'freeXu_lvq' LVQ codebook set.*

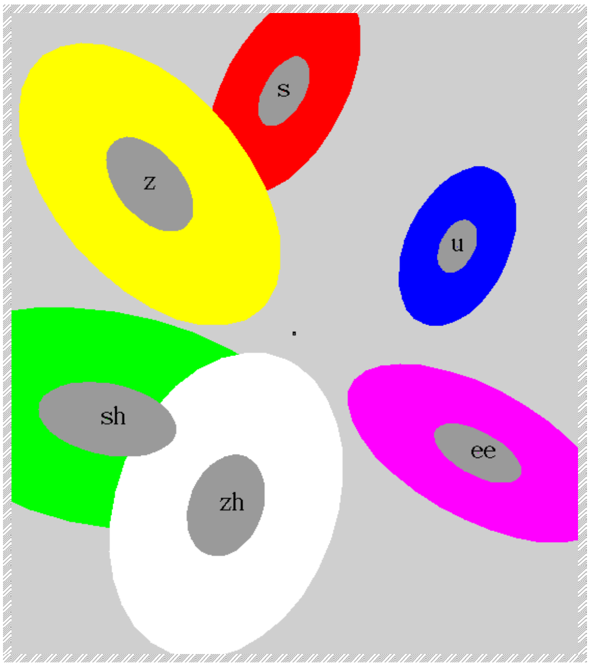


Figure Appendices-53 *Map representation of 'freeXu_lvq' phone classes as ellipses*

File name	*'iso4vo2fr'*					
Description	Normal English male adults map with four vowels and two sibilant fricative targets.					
Subjects analysed	8 normal English male adults					
Data analysed	Vowels and consonants in isolation					
Classes	**Train set**	**Test set**				
/a/	1200	300				
/i/	1200	300				
/o/	1200	300				
/u/	1200	300				
/s/	1200	300				
/z/	1200	300				
Total	**7200**	**1800**				
Net type	**Train set**	**Test set**	**Accuracy**	**Inp**	**Out**	**Hid**
SMANN	0.0004	0.0007	MSSE	9	2	16
PPANN	0.0002	0.0017	Entropy	9	6	16
	100 %	99.89 %	Classify			

Table Appendices-22 *Map configuration for 'iso4vo2fr' file*

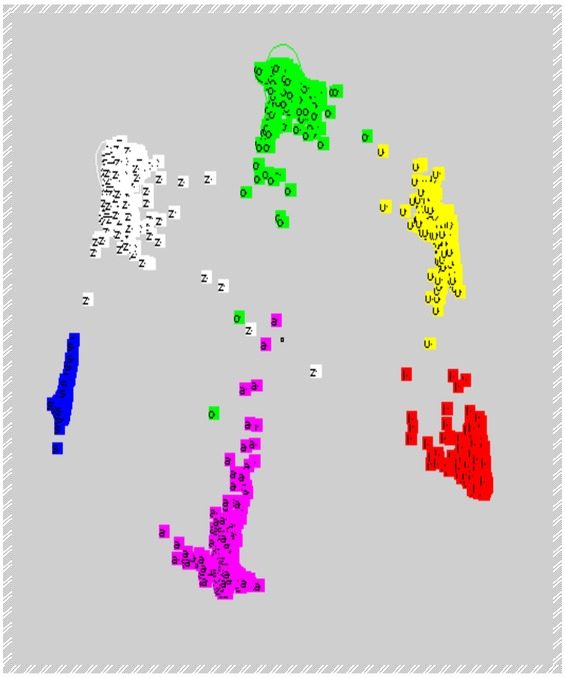


Figure Appendices-54 *Mapping of 'iso4vo2fr' testing data set.*

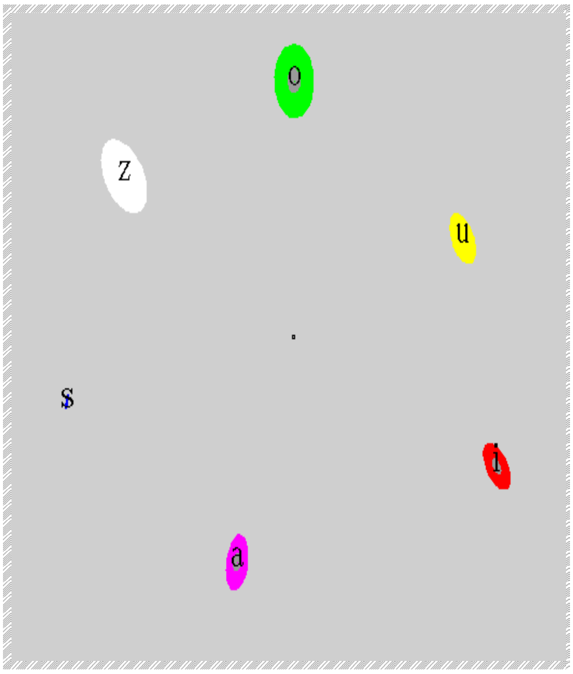


Figure Appendices-55 *Map representation of 'iso4vo2fr' phone classes as ellipses.*

File name	*'iso_s_sh_ee'*					
Description	Normal English male adults map with only one vowel and two sibilant fricative targets.					
Subjects analysed	8 normal English male adults					
Data analysed	Vowels and consonants in isolation					
Classes	**Train set**	**Test set**				
/a/	1600	400				
/i/	1600	400				
/o/	1600	283				
Total	**4800**	**1083**				
Net type	**Train set**	**Test set**	**Accuracy**	**Inp**	**Out**	**Hid**
SMANN	0.0001	0.0001	MSSE	9	2	16
PPANN	0.0010	0.0010	Entropy	9	3	4
	100 %	100 %	Classify			

Table Appendices-23 *Map configuration for 'iso_s_sh_ee' file*

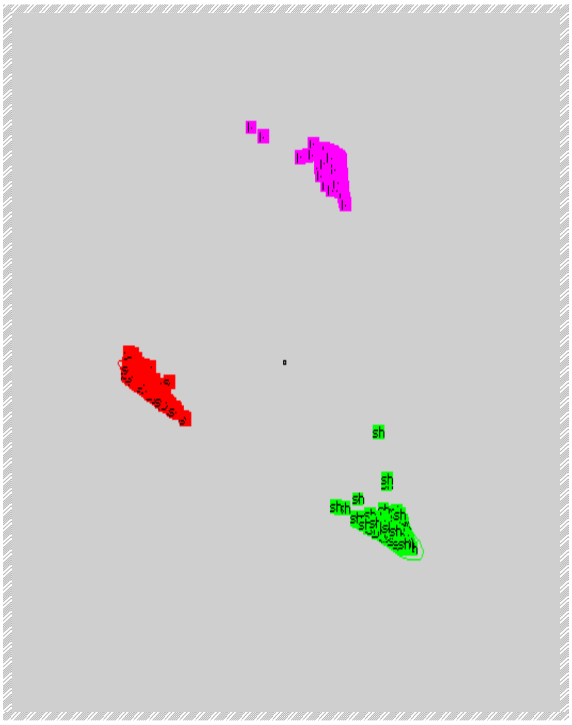


Figure Appendices-56 *Mapping of 'iso_sh_i' LVQ codebook set.*

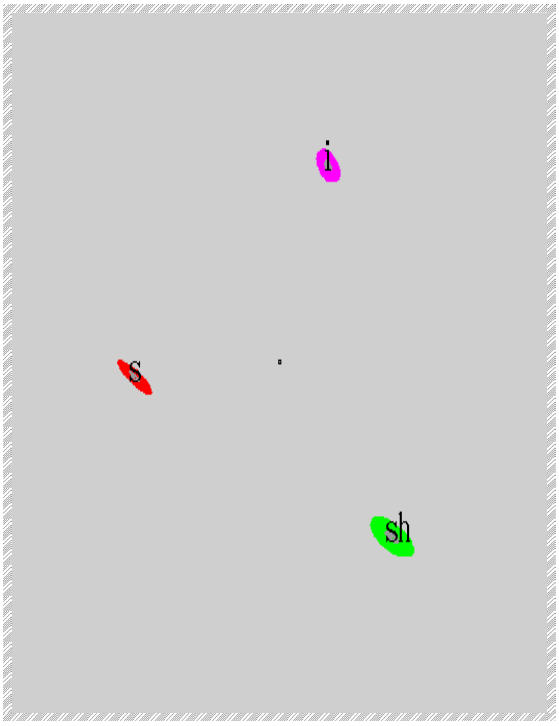


Figure Appendices-57 *Map representation of 'iso_s_sh_i' phone classes as ellipses.*

VIII ***Training sessions schedule***

Session	**Date**	**Stage**	**Clients**
Baseline Recordings	(13/01/99)	1st	ChildB, ChildA, ChildC.
Aural Tests	(20/01/99)	1st	ChildB, ChildA, ChildC.
OLT demonstration and tests	(27/01/99)	1st	ChildB, ChildA, ChildC.
Session 1	(09/02/99)	2nd - Normal maps	ChildB, ChildA, ChildC.
Session 2	(10/02/99)	2nd - Normal maps	ChildB, ChildA, ChildC.
Session 3	(16/02/99)	2nd - Maps with s1	ChildB, ChildA.
Session 4	(17/02/99)	2nd - Maps with s1, s2	ChildB, ChildA.
Session 5	(23/02/99)	2nd - Maps with s1, s2	ChildB, ChildA.
	(09/03/99)	2nd - Map with s1	ChildC.
Post therapy review	(30/03/99)	3rd	ChildB, ChildA, ChildC.

IX Application papers to ethical committee

Contains :

1. The ethics protocol a report to describe to the committee the efficacy study for OLT.

2. The parents information sheet to describe to the parents several issues concerning the treatment of their children with OLT.

3. The child's information sheet to describe to the children how the therapy with OLT will look like.

4. The ethics committee final approval to carry on with the efficacy study for OLT.

X Parents assessment questionnaire for OLT

Contains :

questions made to the parents of the children concerning the assessment of the involvement of OLT in the therapy sessions of their children.

XI Speech and language therapy report

Contains :

The report of the therapist involved in the speech training of the children with the OLT method. She presents the clinical background of the clients, the description of the therapy focusing on how she used OLT during the speech training sessions, and finally an evaluation of OLT. A list of transcriptions is also provided by her to compare the results before and after therapy with OLT.

XII Request form for recordings at schools

Contains :

A letter that was sent to several heads of schools in Sheffield area to ask for permission to make recordings of children that attend their school. A list of children is also provided together with the addresses of the schools involved in the experiments.

Appendix XII – Request Form for Recordings at Schools

Subject ID	Gender	Age	Where
01	Female	7	Colleague's child
02	Male	9	Colleague's child
03	Male	7	Sacred Heart School
04	Male	6	Sacred Heart School
05	Male	7	Sacred Heart School
06	Male	7	Sacred Heart School
07	Male	7	Sacred Heart School
08	Female	6	Sacred Heart School
09	Female	7	Sacred Heart School
10	Female	6	Sacred Heart School
11	Female	7	Sacred Heart School
12	Male	7	Springfield School
13	Female	6	Springfield School
14	Female	8	Springfield School
15	Male	8	Springfield School
16	Female	8	Springfield School
17	Female	8	Springfield School
18	Male	6	Springfield School

Sacred Heart School RC Junior School	**Springfield Infant & Junior School**
Head : Mrs M.A. Bowers	Head : Ms P. Torey
Ripley Street, Sheffield, S6 2NU.	Broomspring Lane, Sheffield, S10 2FA.