Proceedings of the Institute of Acoustics

OPTICAL LOGO-THERAPY - (OLT)
A computer based speech training system for the visualization of articulation using connectionist techniques.

A. Hatzis (1), P.D. Green(1), S. Howard(2)

(1) University of Sheffield, Department of Computer Science
(2) University of Sheffield, Department of Human Communication Science

SUMMARY

Computer-based speech training systems (CBST) are used to provide feedback to subjects with problems in speech production such as phonation, articulation and intonation. The aims of CBST are to display visually one or more attributes of an utterance to assist the speech and language therapist in the assessment of the subject's production or to motivate the subject by providing feedback to reinforce the learning of speech skills.

This paper describes a CBST system called OLT (Optical Logo-Therapy). OLT provides visual feedback to the subject by representing an utterance as a trajectory on a 2D 'phonetic map'. The map is created by Sammon mapping, a technique which defines, for a given set of N-D vectors, a non-linear projection into a 2-D space. The data set for which a Sammon map is created is obtained by Kohonen's learning vector quantisation from a training set of utterances from normal speakers.

The path plotted on the Sammon map for a new utterance provides a qualitative comparison against a standard of speech performance. OLT also contains tools to make quantitative comparison and display the results.

OLT can be specialised to deal with particular speech training problems. Here results are presented for an experiment based on the pronunciation of the sibilant fricatives /ʃ, s, z and zh/. Comparisons are given between a subject with difficulty in articulating these sounds and normal speakers.

1. COMPUTER BASED SPEECH TRAINING AIDS

Some of the critical questions associated with CBST are :

- What kind of feedback is provided?
- How is the speech production of the subject evaluated?
- What guidelines for error correction does the system provide for the therapist and the subject?
- Is training limited to the phone level or is it expanded to the word and continuous speech levels?
- Does the CBST eventually improve the learners' production?
- Does the CBST adjust its training targets as the subject improves her/his speech production ?
- What is the role of the speech therapist?
- How does the subject respond to the feedback provided?

CBST systems can be divided into electro-physiological, articulatory movement and acoustic, depending on the basis of the measurements they use [1]. One of the most developed physiological methods is electropalatography (EPG), which involves placing an artificial palate into the mouth to record tongue-palate contact during speech [3]. EPG has been shown to reveal defects of articulation and is proven to be a useful tool for the speech therapist. However :

- EPG can only be used to investigate phones for which tongue-palate contact occurs.
- In many cases the tongue-palate proximity is not revealed.
- It cannot be applied unless an artificial palate is build for each subject.
- It is an invasive technique.

On the other hand, acoustic systems like Speech Viewer [11], VSA [5], ISTRA [2] and HARP [10] are user friendly and attempt to motify the speech training. Almost all of them are based on automatic speech recognition techniques. Visual feedback is provided in all cases using animation and various ways of rewarding the subject, for instance in games based on recognition scores. Sometimes, however, the targets the subject should reach are too far away from her/his current attempts for recognition score feedback to be meaningful. Most of these packages attempt to deal with all aspects of speech production and sometimes cannot be specialised to treat different groups of users and special speech disorder cases. Furthermore, they do not present a visualisation of the speech events and therefore cannot show where and how a pronunciation has gone wrong. Most importantly, many of these recognition-based CBST systems do not consider the training of a user at the word level or in connected speech, and so cannot train for coarticulation effects.

2. VISUALISING PHONETIC SPACE

Some of the previously referenced systems and other recent ones [8], [12], attempt to visualise the phonetic space, or part of it, in two dimensions. The aim is to project an articulation on to such a map so that the subject and the therapist can see where an error occurs and how an articulation differs from normal. This visual representation of speech can be turned into a computer game using animation techniques.

Kohonen used self-organising maps (SOMs) to visualise phonetic space in his 'phonetic typewriter' [6]. SOMs map patterns in their input space to points in a suitably-chosen output space, usually a two-dimensional grid, preserving topological proximity. By relating output grid activations in a trained net to the phones being spoken , Kohonen effectively reduced the multi-dimensional features of acoustic-phonetics to two dimensions and provided a visual mapping of their similarities. By plotting the trajectory of maximum activation on the map as an utterance is made one can view the path through the phonetic space. Kohonen also mentioned the potential use of such a method in speech therapy.

SOMS were used for speech training in Reynolds and Tarassenko's 'visual ear' [9], which addressed problems related to the smoothing of the trajectory and the provision of a degree of temporal variance. These authors demonstrated the viability of the method in learning pronunciation. Kohonen's team has explored the coarticulation phenomena present in the production of a vowel following a fricative using SOMs [7].

Hatzis developed a graphical user interface called VAHISOM for experimentation and demonstration of SOM phonetic maps [4]. VAHISOM provides displays and menus which allow the clustering on a SOM to be visualised and trajectories to be followed. The current work follows from VAHISOM and attempts to provide better display tools than that package. Recently, Nagayama et. al. have explored the application of another type of neural-

network to phonetically decode speech and represent it visually. Their research is based on the non-linear mapping technique called Sammon mapping [13]. In the work reported here we attempt to combine SOM and Sammon mapping.

3. OPTICAL LOGO -THERAPY (OLT)

I. The interface and its functionality

OLT is at present implemented in C++ and runs on Unix machines. It deploys three windows, the map area, the control panel and the samples pool [Fig 1,4]. In the map area we display a 2D Sammon map and can superimpose a speech trajectory. Each node on the map is associated with a label (in this case a phone label), and different phones have different colours. The arrows on the trajectory indicate the time-course of the utterance. The displayed size of a node on the map can vary depending on how many hits it has received through the whole duration of an utterance. This provides a way of examining the quasi-steady and transient states of the utterance to provide visual

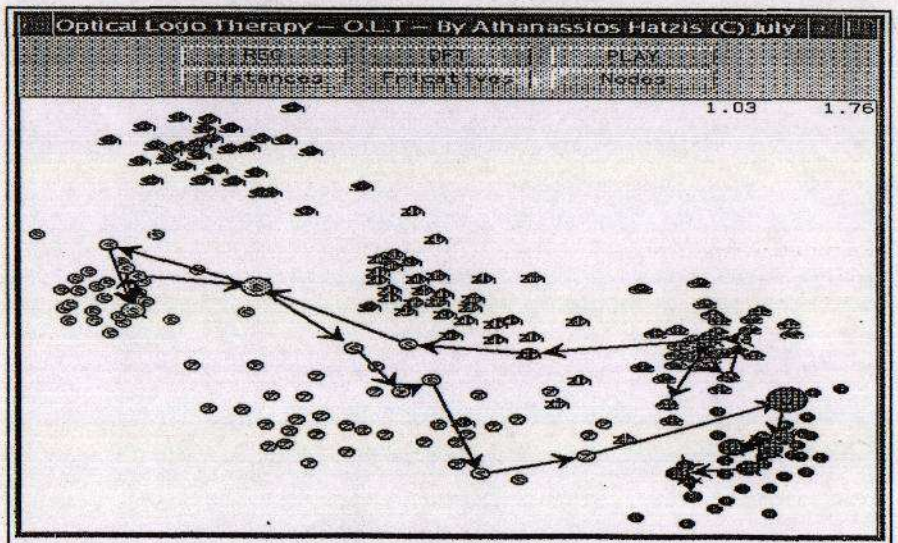

Figure 1 : OLT - The main map area and the trajectory 'i.ʃ.u'

feedback about coarticulation. More than one speech trajectory can be displayed at the same time on the map with different colors and styles and width so that utterances can be easily compared [Fig . 6]

Six buttons are placed on top of the map window:

OPTICAL LOGO-THERAPY - (OLT)

- REC allows a new utterance to be recorded and stored, with the results displayed on the map. Record and display does not, as yet, work in real-time - there is a short delay before the trajectory appears on the map.
- PLAY plays back an utterance through a speaker (either the last one recorded or one selected from a pool) and displays its trajectory.
- DFT computes and displays the DFT spectrogram of the utterance last played, using software in the OGI tools package [Fig. 5]

The lower row of buttons provide various statistical information and displays about the current utterance:

- DISTANCES brings up a display [Fig. 3] which indicates the closest nodes on the map through the time-course of the utterance and the distance between the incoming data vector and the best-scoring node. This is meant to provide a visualisation of the course of coarticulation

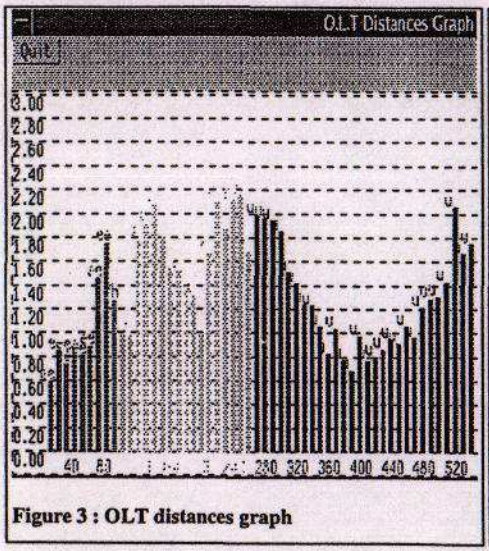

Figure 3 : OLT distances graph

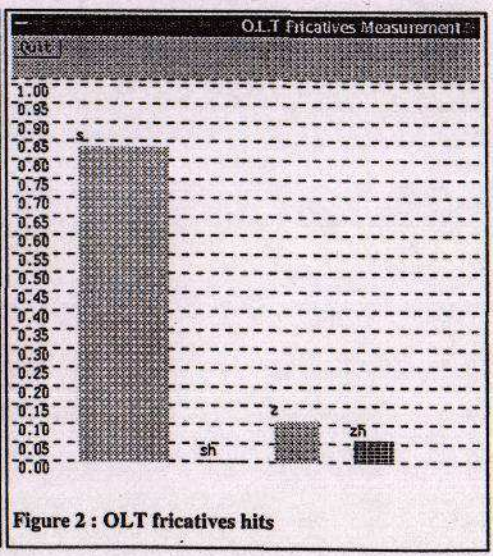

Figure 2 : OLT fricatives hits

- FRICATIVES was added for the current investigation and brings up a display of fricative quality [Fig 2], in the form of a bar chart showing hits for the four fricatives of interest.
- NODES gives access to detailed information related to the nodes on the SAMMON map.

The control panel [Fig. 4] provides tools which handle the events which are displayed in the map area. For instance there are configurable attributes for the drawing of the speech trajectory such as color, style, and width. These are used to portray clearly two or more speech trajectories on the same map. The user can also set the duration of the recording in msec, recording level and playback level. Utterances already stored can be recalled, replayed and displayed on the map by selecting them from the sample pool. The utterances which are in the pool

are defined by a directory mask, so that for instance all the attempts by a single speaker, or all the utterances of a particular phone, can be readily accessed.

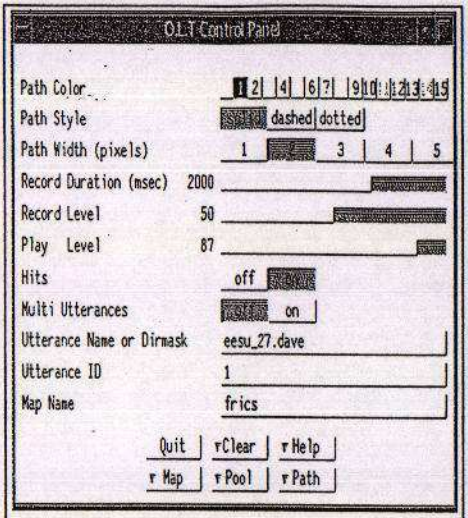

Figure 4 : The OLT control panel

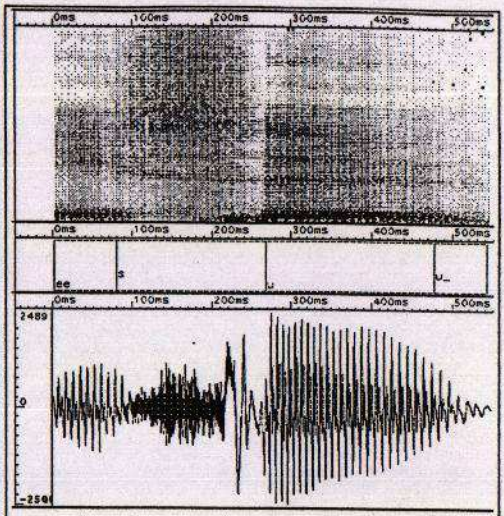

Figure 5 : Segmentation and labelling of an utterance

II. How OLT works

The input to the OLT system is a series of 9D vectors - 8 Mel Frequency Cepstral Coefficients together with overall energy. The frame rate is 10msec. Kohonen's Learning Vector Quantisation [15] is used to obtain a set of feature vectors representative of the clusters in the input data. The Sammon mapping technique is then applied to reduce the 9D reference vectors to points in a 2D space. Sammon mapping is a non-linear projection based on fitting N points in the 2D space such that their inter-point distances approximate the corresponding inter-point distances in the 9D space. The trajectory of an utterance over a Sammon map is obtained by finding, for each vector, the closest reference vector on the map. The sequence of these 'winners' defines the trajectory.

4.EXPERIMENTS

We report on an (unfinished) experiment using OLT to assist in therapy for a subject who has difficulty in pronouncing the sibilant fricatives /ʃ, s, z and zh/. This is a direct application of the system to a specific clinical case and shows how a map appropriate to the case can be constructed.

The speech impaired subject is an adult English male. He was selected because a perceptual analysis of his speech suggested that all of the target sibilant fricatives /ʃ, s, z and zh/ were produced laterally, rather than centrally. That is to say, the perceptual impression was of an airsteam which escaped over one or both lateral margins of the tongue, rather than centrally along a median lingual groove. One of the effects of this misarticulation is to blur the

perceptual distinction between the alveolar and post alveolar targets. Thus the subject's misarticulated /s/ is perceptually indistinguishable from his misarticulation of /ʃ/, even though he may be making consistently different lingual gestures in an attempt to distinguish them [15].

We constructed a map for fricative utterances from six normal speakers, also male, adult and English. To provide a fixed context for the fricative articulation, the sounds spoken were VCV utterances of the form /i X u/, where X is /s, ʃ, z or zh/. Each of the normal speakers recorded 11 repetitions of these sounds, 44 utterances in all. These utterances were segmented manually and labelled as shown in [Fig. 5]. The data from 4 of the normal speakers was used to train the map and the rest was used to test the accuracy on classification. A test vector is classified correctly, if its label matches the one of the closest reference vector. Based on that the overall success rate for all the testing input was 84.53%

The trajectories observed on the map for the test speakers were sufficiently close to those for the training speakers to lead us to hope that the fricative map provides a useful backdrop for this training task.

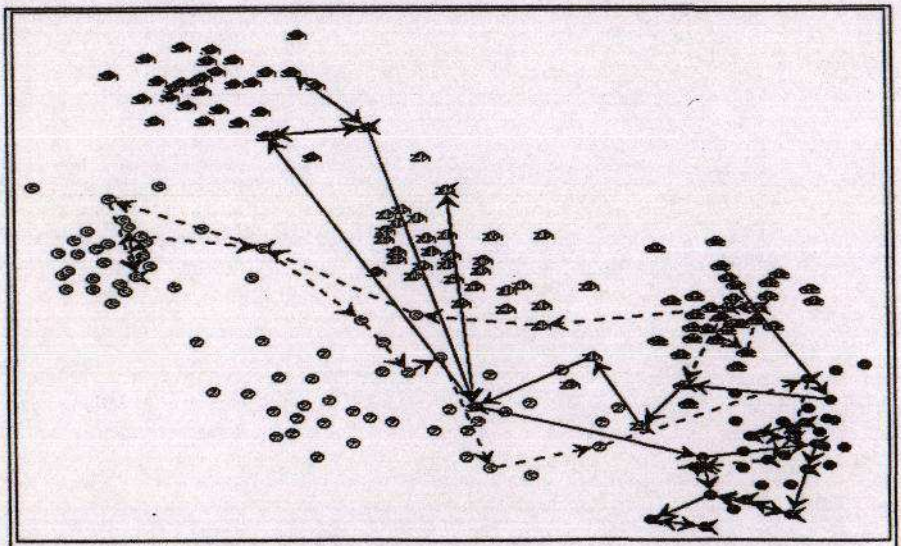

Figure 6 : A comparison of a speech impaired subject (solid line), with a normal speaker (dashed line) pronouncing the utterance 'i.s.u'.

The speech-impaired subject shows clear abnormalities in his vowel-fricative-vowel trajectories, which are consistent with his 'lateralising' problem. [Fig. 6] shows a normal trajectory for 'i.s.u' and an abnormal trajectory, which hits the [ʃ] cluster rather than the [s] cluster. The speech impaired subject spoke at a normal rate. However when he articulates the same utterance slowly, the quality of his speech significantly deteriorates and he cannot not reach the centre of any fricative area on the map [Figs. 7,8].

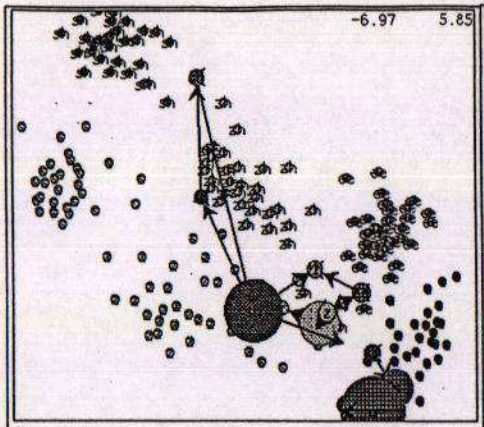

Figure 7 : Abnormal 'i.s.u' - slow rate of speech

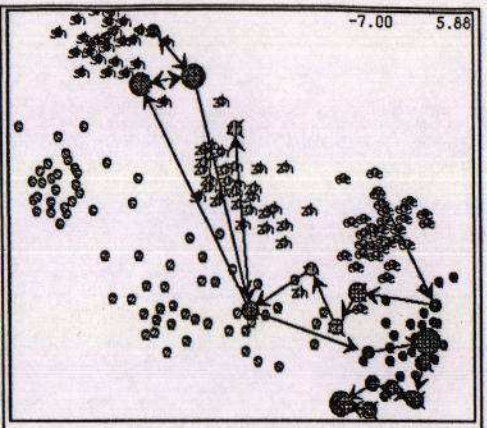

Figure 8 : Abnormal 'i.s.u' - normal rate of speech

5. FUTURE PLANS

Work on OLT is still at an early stage. The eventual aim is to provide a toolkit for studying the articulation of a subject and providing the subject with visual feedback to help improve her/his pronunciation. We have demonstrated the creation of this visual feedback in the form of a trajectory in 2D 'phonetic space', where the display can be tailored to the needs of a particular subject or subject-group.

Much work remains to be done. We intend to address:

- Adding a final supervised learning stage, as suggested by [8] to provide a transformation directly from the input vectors to Sammon map coordinates. This should produce smoother trajectories than the existing technique of finding the nearest reference vector.
- Automating the segmentation and labelling stage used in training using aligment with an ASR device.
- Displaying map coordinates in real time, as the subject speaks.
- Adding animation to improve the display of an utterance trajectory.
- Development of a library of pre-prepared maps for common problems in speech therapy and for different subject groups.
- Maps which adjust themselves with time as the subject's articulation improves.

ACKNOWLEDGEMENTS

A. Hatzis is supported at the University of Sheffield by a Barnsley College studentship.

REFERENCES

[1] D. Kewley-Port, C.S. Watson. Computer Assisted Speech Training : Practical Considerations. In A.Syrdal, R. Bennett & S.Greenspan (Eds.). Applied Speech Technology. Bocal Raton: CRC Press, pages 565-582, 1995

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

[3] WJ. Hardcastle, FE. Gibbon, W.Jones. Visual display of tongue palate contact: Electropalatography in the assessment and remediation of speech disorders. Br J Disorders of Commun. 1991; 26: 41-74

[4] A. Hatzis. Visualisation of the articulation for the hearing impaired with self organising maps (VAHISOM). Unpublished M.Sc. thesis, The University of Edinburgh, Dept. of Artificial Intelligence, September 1995

[5] N.Arends, D.J. Povel, S.Michielsen, J. Claassen, and I. Feiter. An evaluation of the visual speech apparatus. Speech Communication, 10:405-414, 1991.

[6] T. Kohonen. the neural phonetic typewriter. IEEE Computer, pages 11-22, March 1988

[7] L. Leinonen, R. Mujunen, J. Kangas, and K. Torkkola. Acoustic Pattern Recognition of Fricative - Vowel Coarticulation by the Self-Organising Map. Folia Phoniatrica, 45:173-181, 1993

[8] I. Nagayama, N. Akamatsu, T. Yoshino. Phonetic Visualisation for Speech Training System by Using Neural Network. In Proc. International Conference on spoken Language Processing (ICSLP94), pages 2027-2030, 1994

[9] J. Reynolds and L. Tarassenko. Learning Pronunciation with the Visual Ear. Neural Computing and Applications, pages 169-175, 1993.

[10] E. Rooney, M. Jack, J. Lefevre and A. Sutherland. HARP-a speech training aid for the hearing impaired. In 2nd TIDE Congress, la Villette, Paris, April 1995.

[11] J. Ryalls. Comparison of two computerized speech training systems: Speech Viewer and ISTRA. Journal of Speech-Language Pathology and Audiology, 13(3):53-56, 1989

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

[13] J.W. Sammon. A Nonlinear Mapping for Data Structure Analysis. IEEE Trans. on Computer, C-18, 5, pages 401-409, 1969.

[14] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, and Kari Torkkola. LVQ_PAK: The Learning Vector Quantization Program Package. Technical report A30, Helsinki University, Faculty of Information Technology, Laboratory of Computer and Information Science, 1996

[15] S. Howard. Spontaneous phonetic reorganisation following articulation therapy: An electropalatographic study. In R. Aulanko & A-M. Korpijaakko-Huuhka (Eds) Proceedings of the 3rd Congress of the International Clinical Linguistics & Phonetics Association, 67-74, Helsinki: Publications of the Department of Phonetics, University of Helsinki, 1994