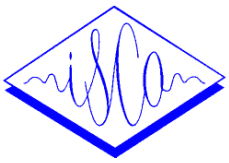

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

A SPEECH THERAPY WORKSTATION FOR THE ASSESSMENT OF SEGMENTAL QUALITY: VOICELESS FRICATIVES.

A.A.Wrench[1], M.S. Jackson[2], M.A. Jack[1], D.S. Soutar[2], A.G. Robertson[3] and J. MacKenzie[4], J. Laver[1]

1 Centre for Speech Technology Research, University of Edinburgh, 80 South Bridge, Edinburgh, UK.
2 Plastic Surgery Unit, Canniesburn Hospital, Bearsden, Glasgow, UK.
3 Beatson Oncology Centre, Western Infirmary, Glasgow, UK.
4 Department of Speech Pathology and Therapy, Queen Margaret College, Edinburgh, UK.

ABSTRACT

A speech therapy workstation is under development for the purpose of supervised rehabilitation of oral cancer patients. The design of this workstation differs from that of most others of this type in three ways. Firstly, the software is designed to address the problems of this specific patient group. Secondly, patients and therapists are involved throughout the design process at the earliest opportunity. Thirdly, the project is based on advances in automated acoustic phonetic analysis which overcome many of the problems associated with formant analysis of voiceless fricatives and in so doing provide intuitive visual feedback for this category comparable to the F1/F2 plots for vowels. This paper discusses the constraints placed on the design of such a workstation, the solutions and the acceptability of the resulting system.

Keywords: *Speech therapy, visual aid, objective speech analysis, glossectomy speech.*

1. BACKGROUND

It is only in the last decade that the computational power/cost ratio of computer resources has risen to the level where an affordable system can present visual speech patterns while they are being spoken. Over this period speech technology researchers have focussed renewed interest in the field of speech rehabilitation and training. Initially this has been a process of direct transfer of techniques from the fields of speech analysis and recognition.

The segmental analysis modules in IBM's Speech Viewer II™ and the ISTRA system [4] are based on the spectral template matching procedure commonly employed for speech recognition tasks. The CASTT system [7] and SAS system [1] include a vocal tract reconstruction module based on Linear Predictive Analysis to display an estimate of the vocal tract configuration for vowels. The CASTT system has additionally a fricative monitor, spectrogram and phoneme plotter providing further visual feedback of segmental quality. A wide range of speech analysis probes and sensors have been developed. The CISTA system [9] and Kay CSL™ system both use multi-sensor input to enhance the robustness of segmental analysis.

More recently the extraction formant structure for vowels has been realised in several functional systems. The CISTA system uses an F1/F2 plot derived from a centre of gravity measure to provide visual feedback of vowel production while the CSL system and SPELL pronunciation aid [6] use LPC peak tracking algorithms for the same purpose. The VSA system [5] provides an approximate F1/F2 plot by performing discriminant analysis on filterbank input. The vowel articulation training aid [10] takes this a stage further by applying non-linear transformation to the filterbank prior to linear transformation and providing enhanced vowel separation in two dimensions. Detailed acoustical analysis of consonants has, however, been slower to develop.

2. INTRODUCTION

Designing the workstation specifically for the requirements of intra-oral cancer patients allows research to concentrate on those issues of importance to this patient group providing a system tailored to their requirements. Early discussions between the speech therapist, speech scientists and design engineers isolated the areas of greatest need. These were found to lie mostly in the production of consonants. Current systems that provide visual feedback of consonant performance use spectral templates. A template matching approach provides a guage that indicates the degree to which the sound matches the target template on a scale of spectral similarity. This approach, however, provides indistinct

feedback when a user attempts to adjust articulation to improve the match. The work reported here is based on the design of a method of formant analysis applicable to consonants which can provide more helpful visual reactive feedback choosing an approach similar to the F1/F2 plots commonly implemented for vowels.

3. DESIGN CONSIDERATIONS

During a speech therapy session patients become apidly fatigued. Short sessions of about 20 minutes are sometimes as much as can be managed in each of the first few sessions after treatment. Pre-operatively, patients are sometimes only available for even shorter periods. This time constraint implies that statistical modelling of individual patient's speech cannot be relied upon as a method of analysis because sufficient training data is unlikely to be available from any given subject.

Patients do not have time to become familiar with the workstation so the user interface must be robust and intuitive. The protocol must be simple. In contrast, the speech therapist using the workstation has the opportunity to become familiar with its operation. However with therapy session time at a premium it is necessary to ensure that as little time as possible is taken up by configuring the system for each patient and that the flow of the session is not unduly disrupted by periods spent waiting for or managing the system. The workstation is therefore designed to require only a few seconds to calibrate for the patient and recording environment prior to each session calibration.

The phonetic classes that cause the most problems for oral-cancer patients are those which require a high degree of lingual control. Working with the intra-oral cancer patient group narrows the field of relevant speech errors from the six hundred that a therapist might observe in common practice [2] down to a subset primarily relating to lingual (tongue) mobility. In addition there are a number of errors related to the changes in oral topogaphy caused by reconstructive surgery which are peculiar to this patient group. To date no extensive survey of the articulatory errors has been carried out but a list of phonetic categories most affected by intra-oral treatment, in order of their combined importance to speech quality and susceptibility to therapy, can be specified:

/l/
/t/,/k/,/d/,/g/
/s/,/ʃ/,/z/,/ʒ/,/θ/,/ð/

Table 1. *Important phonetic classes requiring speech therapy for patients who have undergone surgery and/or radiotherapy of the oral cavity/tongue.*

When considering the priority of segmental categories for inclusion in the workstation, the issue of automatic segmentation determined the choice. Automatic isolation of the sound of interest is an important issue in relation to ease of use. From the list in Table 1.the voiceless fricative class is the least complex to isolate from a continuous utterance. This can be done by applying voiced/ unvoiced/ silence detection [8] and choosing an appropriate carrier utterance. Furthermore, research in this category can contribute to the design of the more complex voiceless stop category for which fricative analysis will form one component. More sophisticated automatic segmentation procedures are being considered for the future to facilitate the analysis of the remaining acoustic segments listed in table 1.

4. MEASURING THE QUALITY OF /s/

Speech quality has essentially two components: intelligibility and acceptability. The current design relies on professional observations made by the speech therapist regarding the key factors contributing to speech quality of /s/ after treatment. For the purpose of visual feedback, the basic requirements of the module are to distinguish and quantify tongue retraction; over-compensation of tongue position in the forward direction and lateralisation while taking into account the effects of lip rounding. For the purpose of objective quality assessment the degree of nasalisation should also be measured. The study of acoustic phonetics leads to an understanding of spectral structure of voiceless fricatives. Primary, secondary and even tertiary fricative excitation sources (here ranked by their relative energies) exist for voiceless sibilant fricatives. The primary frication results from airflow directed onto the teeth producing turbulence at this obstacle. Secondary frication is produced by the constriction through the grooved channel between the tongue and roof of the mouth. A tertiary source is possible from a constriction further down the vocal tract; for example at the glottis. As a result of the source(s) being part way along the vocal tract, the spectral structure of voiceless fricatives may have 0, 1 or 2 formants due to cavities beyond the source and 0, 1 or 2 anti-formants due to acoustic coupling with cavities before the source. These formants and anti-formants vary in frequency and amplitude in a semi-dependent manner according to lip, tongue and jaw position and dentition. Ideally therefore, the analysis process should be able to detect the number, frequency and amplitude of formants and anti-formants.

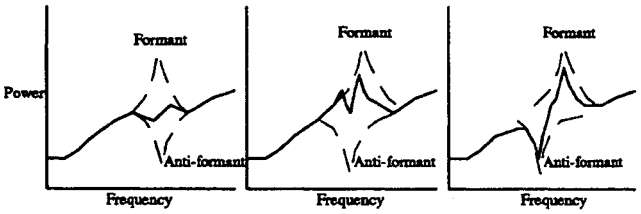

Figure 1. *Three spectral distributions resulting from the combination of a formant and an anti-formant*

The anti-formants may have one of three possible spectral manifestations. Firstly an anti-formant may cancel a formant if it has the same frequency. Secondly it may split a formant

resulting in two spectral peaks. Thirdly, it may result in a notch in the power spectrum, Figure 1.For mono-lateral and bi-lateral fricatives it is possible for the primary and secondary sources to be reversed with the primary source being at the point(s) where the seal is broken. However, the spectral structure is still defined by the parameters outlined above.

5. TECHNICAL DESIGN OF THE /s/ MODULE

The existence of formants and anti-formants (poles and zeros) in the spectrum suggests the application of an Autoregressive Moving Average (ARMA) model. However the ARMA model is unsuitable for three reasons. Firstly, on a practical level, there is no known method of solving the non-linear equations associated with the ARMA model: An approximate iterative or decomposition method must be used. Secondly, it is not known *a priori* how many poles and zeros there will be for an unknown fricative. In addition, there may be more than one source or a distributed source for which this model does not apply. A precise mathematical solution of this kind is therefore not practical. The system described here is designed to identify the peaks and troughs of the spectrum and classify sounds on the basis of these patterns. Often these features will correspond to the formants and anti-formants of the acoustic speech signal.

Identification of a single spectral peak may be done by centroid analysis (also known as centre of gravity or first spectral moment[3]. In order to identify more than one peak, methods are required for the determination of multiple centroids of a single distribution simultaneously. The fricative analysis module reported here implements novel dual centroid analysis which fulfils the criteria of determining the number, frequency and band energy of up to two spectral peaks, Figure 2. Furthermore, it is possible to identify up to two spectral troughs by applying the same analysis to the reciprocal power spectrum.

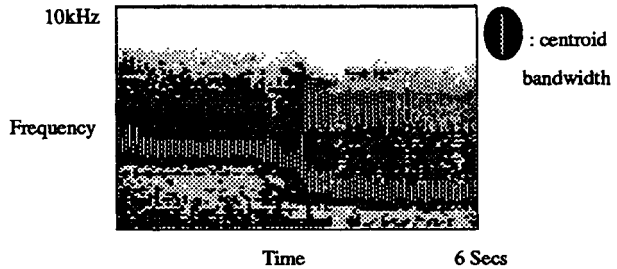

Figure 2. *Spectrogram showing dual centroid analysis of /s/ in transition with tongue retracting and slight lip rounding.*

The voiced/unvoiced/silence detection; 256 point fast Fourier transform; Bark scaling; dual centroid analysis and graphical display are performed on a 33MHz 486DX processor every 100ms providing instant visual feedback. This is a satisfactory rate of visual feedback for sustained segments.

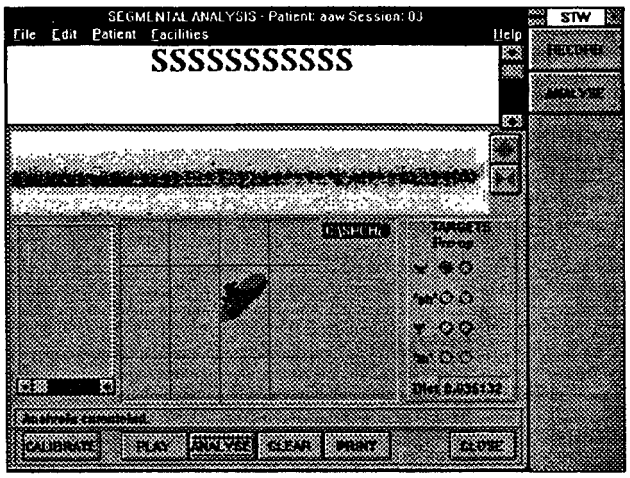

Figure 3 *Screen image for system.*

6. DESCRIPTION OF DISPLAY FOR /s/

The visual feedback display is positioned in the centre of the screen, Figure 3. Bold text prompts are sited above it for the patient to refer to if necessary. The feedback display has two axes. The horizontal axis corresponds to the frequency of the spectral peak with the lowest frequency (Bark scale). The vertical axis corresponds to the frequency of the peak with the highest frequency (Bark scale). Within the six second analysis period an accumulation of coloured dots appear on the screen at a rate of one every 100ms while the fricative is being uttered into the microphone. Periods of silence or voiced speech produce nothing on the screen. If the analysis identifies only one peak then the dots appear somewhere on the diagonal defined by upper peak equals lower peak. Otherwise, if two peaks are found then the dots appear somewhere above this diagonal, Figure 4. Further processing is required to identify and provide visual feedback of fricatives with no spectral peaks.

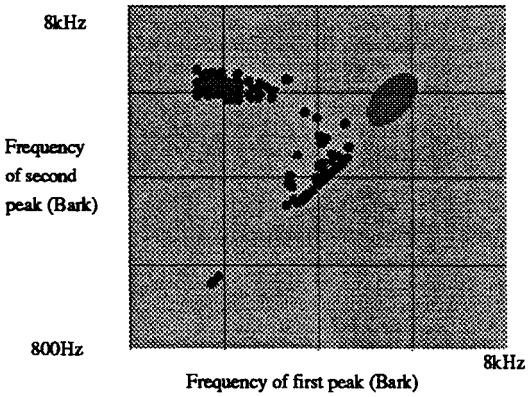

Figure 4. *Visual feedback for the fricative analysed in Fig. 2.*

Targets for the production of /s/ and /ʃ/ can be selected and shown in the background of the display by the therapist using a selection panel to the right of the feedback display. One target at a time may be shown in the background in the form of a dark grey ellipse.

7. CHOOSING A REFERENCE FOR QUALITY

The choice of reference is one of the most important aspects of the design of a functional speech quality assessment system. Two target types were considered. These were: The patients own speech target recorded pre-operatively; and a standardised target based on an average of all patients' speech. The former was chosen since speech varies greatly between patients depending on accent and on the size and shape of a patient's vocal tract and dentition. This target however has been found to have several drawbacks. The speech of patients recorded pre-operatively may not be appropriate reference because in some cases the tumour may be sufficiently advanced to affect speech quality and in others a biopsy may already have been performed. In addition, the target generation process itself is unsatisfactory because it requires careful selection of appropriate segments or alternatively, sufficient segments to form a reliable average. We are currently considering an alternative which would consist of selecting one of a set of pre-determined targets according to a best fit to pre-operative data. In circumstances where a patient's speech has been judged to be affected pre-operatively a more suitable target might be selected by the therapist.

8. HOW THE WORKSTATION IS USED

The speech therapist has the role of carer with respect to the patient's post-operative well-being especially in regard to swallowing and communication The therapist provides support and advice to patients as well as assessing functional speech disorders and providing rehabilitative tuition. It is not always possible in this atmosphere to make sure that a computer system has been set up correctly. Such features as insensitivity to microphone placement; to the amplification level; automatic calibration for background noise and speech level; and helpful warning messages have been found to greatly enhance the systems usability. It is important to consider the psychological impact that the use of a workstation of this kind can have on patients. It is essential that the patients are not discouraged during their rehabilitation. The workstation is therefore employed as a means of reinforcing improvements in speech production and rarely to demonstrate the faults. As an exception to this rule it is sometimes used to demonstrate that a patient has achieved better speech quality in a previous post-operative session and therefore that it is possible for them to do better.

9. SUMMARY

A speech therapy workstation is being developed to aid the rehabilitation of patients undergoing treatment for intra-oral cancer. The workstation is designed to be used in speech therapy sessions to be used to provide simple, robust continuous visual feedback to the patient while they are speaking. The purpose of the module described here is to provide positive feedback for the patient when they pronounce fricatives /s/ and sh/. Off-line (in the patient's absence) it is intended that the workstation be used to assess the quality of voiceless fricatives and in so doing to objectively monitor individual patients progress during the period of rehabilitation. Work is continuing to improve analysis of this segmental category and in the construction of a suitable segmental assessment protocol.

ACKNOWLEDGEMENTS

This project was funded by the British Cancer Research Campaign.

REFERENCES

[1]: *Aguilera, S.; Berrojo, M.A.; Giménez de los Galanes, F.M.; Colás, J.; Macías, J.; Montero J.M.*: Impaired persons facilities based on a multi modality speech processing system. Proc. ESCA Workshop on Speech and Language Technology for Disabled Persons, pp.129-132, 1993.

[2]: *Braeges, J.L.; Houde, R.A.*: Use of speech training aids. Deafness and Communication:Assessment and Training, Ed. D. Sims, G. Walter and R.Whitehead, Published Baltimore, Williams and Wilkins, 1982

[3]: *Fujisak H.;Kunisaki O.*: Analysis, recognition and perception of voiceless fricative consonants in Japanese. Annual Bull. RILP 10, pp. 145-156, 1976.

[4]: Kewley-Port, D.; Watson, C.S.; Elbert, M.; Maki, D.; Reed, D.: The Indiana Speech Training Aid (ISTRA) II: Training curriculum and s elected case studies. Clinical Linguistics and Phonetics, Vol.5, No.1, pp. 1 3-38, 1991.

[5]: *Povel, D.: Arends, N.*: The Visual Speech Apparatus:Theoretical and practical aspects. Speech Communication, Vol.10, pp. 59-80, 1991.

[6]: *Rooney, E.; Vaughan, R; Hiller, S.; Carraro, F.; Laver. J.*: Training vowel pronunciation using a computer-aided teaching system. These proceedings.

[7]: *Watson, C.J.; Andreae, J.H.*: To test and assess the remedial worth of a computer based speech therapy aid. Proc. 4th Australian Conf. Speech Science and Technology, pp 279-284, Dec. 1992.

[8]: *Wrench, A.A.; Jack, M.A.; Laver, J.; Jackson, M.S.; Soutar, D.S.; Robertson, A.G.; MacKenzie, J.*: Objective Speech Quality Assessment in Patients with Intra-Oral Cancers: Voiceless Fricatives, Proc. ICSLP 92,Banff, Vol 2, pp 1071-1074, 1992.

[9]: *Yameda, Y.; Murata, N.; Javkin, H.; Antonanzas-Barroso, N.; Das, A.; Niedzielski, N.; Levitt, H.; Youdelman, K.*: A multi-parameter speech training system. Proc. ESCA Workshop on Speech and Language Technology for Disabled Persons, pp.137-140, 1993.

[10]: *Zahorian S.A.; Venkat S.*: Vowel articulation training aid for the deaf. Proc. IEEE. ICASSP-90, pp. 1121-1124, Apr., 1990.