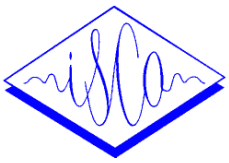


2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

OBJECTIVE SPEECH QUALITY ASSESSMENT IN PATIENTS WITH INTRA-ORAL CANCERS: VOICELESS FRICATIVES

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

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

ABSTRACT

This paper presents an acoustic-phonetic analysis of voiceless fricatives produced by patients undergoing treatment for intra-oral cancer. A method is proposed with the purpose of regularly assessing the quality of such speech before and during rehabilitation. Speech quality is to be measured in terms of accuracy and separation of speech sounds within this class, in comparison with the patient's own pre-operative speech.

1. INTRODUCTION

Background

Between 40 and 50 patients with advanced cancers of the oral cavity are treated each year in the West of Scotland Regional Plastic Surgery Unit at Canniesburn Hospital. Current treatment involves radical excision with immediate reconstruction followed by radical post-operative radiotherapy. Subsequent speech therapy is carried out since the ability of the patient to maintain intelligible speech contributes greatly to post-operative quality of life.

Combined radical treatment has improved the long term survival in patients with advanced tumours so that the likelihood of surviving for at least five years after treatment is, in most cases, better than 60 percent.

Despite these advances, a recent study carried out at Canniesburn Hospital has revealed that 64% of long term survivors experience trouble with speech following treatment [1][2].

Project Goals

The primary goal of this collaborative research into objective and quantitative methods for characterising the speech of patients involved in such oral cavity surgery is to determine what improvements in current practice can be achieved with respect to speech therapy and modification of surgery.

A secondary goal of the project is to evaluate the use of a speech workstation as a rehabilitation aid for use by the speech therapist. Objective speech analysis metrics will be interpreted graphically, enabling the therapist to provide the patients with visual and audible feedback of their speech performance.

Existing Speech Quality Assessment Methods

Existing speech quality assessment methods are largely subjective, such as the Functional Intra-oral Glasgow Scale, FIGS, used as a baseline reference for this research (Figure 1.)

I can chew	
Any food, no difficulty	5
Solid food, with difficulty	4
Semisolid food, no difficulty	3
Semisolid food with difficulty	2
Cannot chew at all	1
I can swallow	
Any food, no difficulty	5
Solid food, with difficulty	4
Semisolid food only	3
Liquids only	2
Cannot swallow at all	1
My speech is	
Clearly understood always	5
Requires repetition sometimes	4
Requires repetition many times	3
Understood by relatives only	2
Unintelligible	1

Figure 1. FIGS Scale

This scale has shown statistically significant reliability between assessors and can be used by doctors, nurses, therapists or dieticians; or as a means of self-assessment by the patient or their relatives. It provides useful information on the progress of a patient through both treatment and post-treatment phases and is currently in use in clinical practice at Canniesburn.

The FIGS scheme operates on a five point scale where a score of 5 represents clearly understood speech: a score of 1 represents unintelligible speech. The FIGS scores are compiled as a longitudinal scale to assess temporal improvements. At Canniesburn patients who score below 3 on this scale are referred for speech therapy treatment.

The assessment of patient's speech using FIGS, although useful, does not identify the source of problems associated with speech disorders nor does it evaluate in scientific terms the quality of patient's speech. The research project described here addresses both of these points.

2. OBJECTIVE ANALYSIS

Speech Data

Patient's speech is recorded digitally in the clinic using a PC-based workstation. Every patient enrolled in the project

10.21437/ICSLP.1992-320

currently records 6 spoken sentences prior to undergoing treatment. The same sentences are then recorded at regular intervals after treatment. The sentences have been designed to be phonetically balanced and contain a range of potentially problematic phonemes. All recordings are made under the supervision of the project speech therapist.

Design Criteria

It is necessary to choose speech distance metrics that are tolerant of patient-specific characteristics such as vocal tract size but sensitive to the position and movement of the articulators (lips, tongue and mandible). Accounting for individual characteristics by modelling each patient incurs a penalty of an enrolment period required by each patient prior to using the workstation. Large amounts of training speech have been ruled out as too exhausting for many patients. The enrolment protocol therefore should involve at most a short calibration utterance before each patient's speech is analysed.

Although the ideal speech rehabilitation pattern would lead to the return of normal mobility and flexibility of the articulators, this is not a reasonable expectation in many cases particularly when substantial surgical excisions have been made. Novel articulations are often required in order to best approximate some sounds. These compensatory articulations may be perceived as good approximates so producing good speech quality. In some cases however, the ability to make each phoneme distinct from all other phonemes and to achieve this consistently is the patient's primary consideration. The measurement must therefore reflect these indicators of speech quality. Firstly the metric should be able to quantify the distance between each post-operative speech token and the patient's own pre-operative target (accuracy measurement). Secondly, the metric must quantify the distance between post-operative tokens corresponding to confusable phonemes spoken by the same patient (separation measurement).

3.CREATING A FEATURE SPACE FOR VOICELESS FRICATIVES

Method

The problem addressed here is the detection of voiceless fricatives within an utterance and their assessment in a manner that permits a qualitative judgement to be made automatically. The chosen method of achieving this is to create a feature space in which accuracy and separation measures can be made.

In operation the system applies signal processing techniques that can automatically analyse selected acoustic phonetic features and provide consistent measurements. This involves two stages. The first stage is to develop a method of identifying the broad nature of each frame of speech; whether it is voiced, unvoiced or non-speech. Once a frame has been assigned to one of these broad classes, it must be further processed to produce features which distinguish between voiceless speech sounds.

In the work to date formant analysis has been used to create such features. These are then combined to create a feature space in which vectors that correspond to perceptually distinct sounds are well separated.

Calibration

All speech recordings are made in the clinic using a head mounted close-talking microphone which is correctly positioned by the speech therapist. The recording level is manually adjustable and may be set at the start of each recording session to compensate for the loudness of patient's speech ensuring a high signal level.

The measurement of relative energy levels is an important component of speech analysis. Therefore, a method of calibration has been implemented to normalise both the absolute noise energy level and the speech/noise energy ratio. This normalisation is consequently applied to all band energy measures used for speech quality assessment.

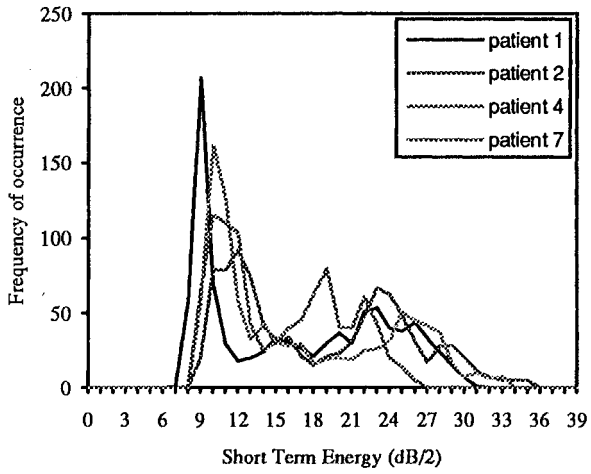


Figure 2.a) Energy distributions of four patient's utterances before normalisation.

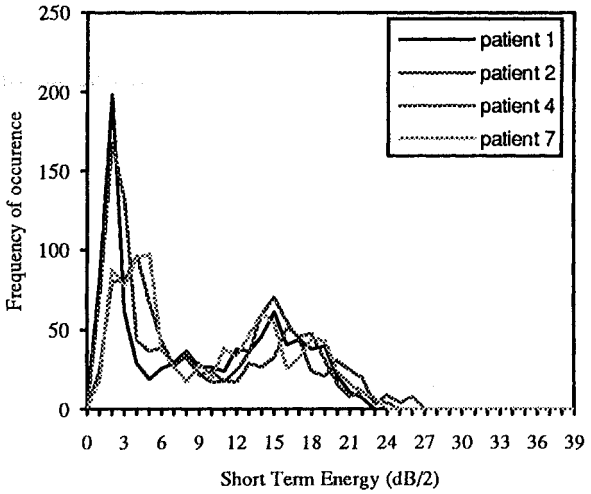


Figure 2.b) Energy distributions after normalisation.

The calibration is performed using a single sentence spoken at the start of each session. Short term energy, measured at regular intervals throughout the utterance, is bracketed into one of 40 levels and accumulated to determine the distribution of signal energy vs. frequency of occurrence (Figure 2a). Bimodal analysis is then performed on the distribution to calculate the mean noise and signal levels.

Consequent energy measures calculated from recordings in the same session are normalised by subtracting the minimum noise level and amplifying the resultant value by the ratio of the mean speech energy to a fixed arbitrary standard energy (eqn. 1).

$$E_{Norm} = (E_{In} - E_{Min.Noise}).E_{Speech}/E_{Fixed} \quad (1)$$

Figure 2b) shows that this method of normalisation succeeds in improving the alignment of speech energy distributions produced by different speakers.

Voiced/Unvoiced/Silence Detection

Analysis is performed on a frame by frame basis. A variable spectral preemphasis factor and a frame energy measure are calculated from zeroth and 1st autocorrelation lag terms (eqn. 2).

$$\text{Preemphasis} = R_1/R_0$$
$$\text{Total Energy} = R_0 \quad (2)$$

These features enable separation of hand labelled frames of patients speech into three clusters: voiced speech (nasals and vowels), unvoiced speech (weak fricatives and strong fricatives) and background noise (figure 3). Two boundaries were calculated using linear discriminant analysis [3] which were then used for a hierarchical frame assignment. First a decision on whether the frame contains speech or not followed by a voiced/unvoiced decision. Voiceless speech frames are selected for further processing.

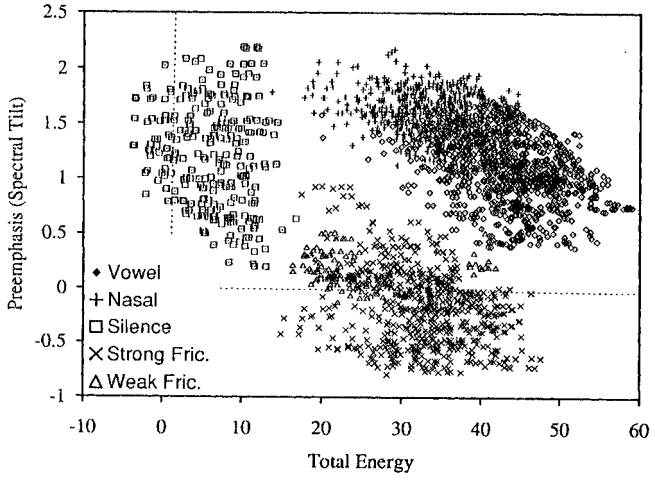


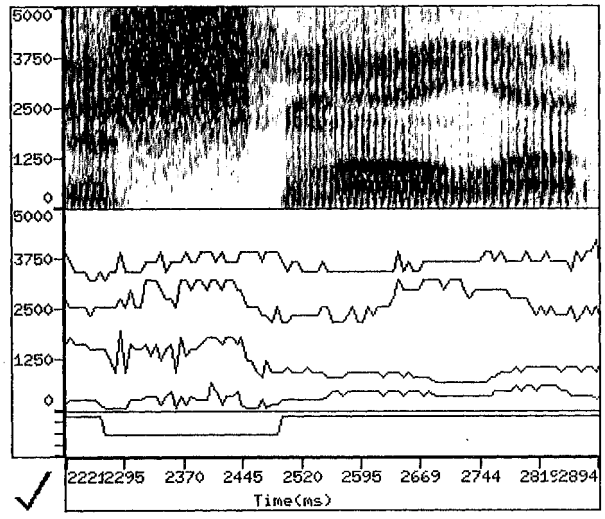
Figure 3. Scatter plot of patient's speech showing separation of voiced, unvoiced and silent data in the dimensions of spectral tilt and energy. Tilt is calculated: if $R_1/R_0 > 0$ then Tilt = $-\log_{10}(1 - R_1/R_0)$: if $R_1/R_0 <= 0$ then Tilt = $\log_{10}(1 + R_1/R_0)$.

Feature Analysis

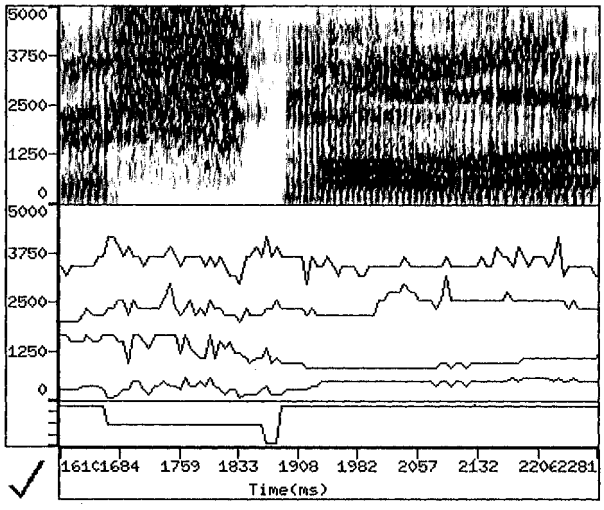
An FFT is performed on each frame of speech. Parameter settings are flexible but a common set-up is :

- 10kHz sample rate
- 12.8ms frame length
- 6.4ms frame shift
- 128 point FFT

Spectra from frames, classified as unvoiced speech, are then Bark scaled before being analysed to identify formants and measure their frequencies and band energies. The algorithm [4] used to do this operates directly on the power spectrum to calculate the centroids of two partitions in the frequency domain. The width and position of these bands within the spectrum are constrained using knowledge of the range of movement of speech formants [5]. The algorithm then splits the spectrum into partitions such that a global error measure is minimised and these bounding constraints are satisfied. The mean and standard deviation of these are calculated and interpreted as formant frequencies and bandwidths. The frequencies and energies of the formants produced by this algorithm and the total frame energy previously calculated currently form the basic feature set for the voiceless fricative class. Figure 5 shows spectrograms of a patient's speech before and after surgery with formant frequencies superimposed and voiced/unvoiced/silence assignment displayed beneath.



a)



b)

Figure 6. Spectrogram of patient's speech "..smaller..." recorded a) pre-operatively b) post-operatively with 3 level display of automatic voiced/unvoiced/silence decision displayed beneath (*top* voiced, *middle* unvoiced, *bottom* silence).

Linear discriminant analysis of the features optimises the separation of voiceless fricatives and creates a feature space in which a standard Euclidean distance metric can be used.

Of the four voiceless fricative phonemes in standard English (/s/,/sh/,/f/,/th/), it is /s/ that provides the most difficulty for patients because the muscular co-ordination required to produce this sound is very high. To make the sound a lateral seal of the tongue with the gums is formed and the tongue is curved to direct the turbulent air flow centrally. Although phonological distinction is made between only four sounds, there are many sounds in this class that are perceived to be distinct. Acoustically these sounds vary according to changes in parameters such as lip rounding and tongue body and tip placement. When considering patients with varying degrees of tissue excised, the variety of sounds increases still further.

Table 1. Voiceless fricative classes

Tongue position	Lip position
grooved alveolar fricative; central airflow	neutral and rounded (GAN, GAR)
grooved alveolar fricative; nasal leakage	neutral and rounded (NAN, NAR)
flat alveolar fricative; broad air flow	neutral and rounded (FAN, FAR)
lateral alveolar fricative ; lateral air flow	neutral and rounded (LAN, LAR)
palato-alveolar fricative;	neutral and rounded (PAN, PAR)
retroflex palato-alveolar fricative	neutral and rounded (RPAN, RPAR)

We consider here the possible acoustic variation of /s/. Linear discriminant training requires speech frames labelled according to linearly separable groups. Deciding what these groups should be proves to be non-trivial. Using the four phonetic classes provides a set that is too course since quality is perceived to vary within these classes. A larger set of twelve articulatorily different classes is proposed (Table 1) which correspond to perceptually distinct variations of /s/. Assigning classes in this way leads to difficulty when hand labelling patient's speech for training because their articulations are not known. We have therefore chosen to use the speech of a group of expert phoneticians to train the feature space, each of whom recorded an example of each sound in isolation. Patient's speech is then assessed within the feature space generated in this way.

4. DISCUSSION

The pre-operative and post-operative recordings of figure 6. were processed and are shown in figure 7 plotted in the 2 most discriminant feature space dimensions. The pre-operative trace has two clusters; the first close to the lip-rounded grooved alveolar class centroid (GAR) and the second close to the nasalised alveolar class centroid (NAN). The post-operative trace also has two clusters; the first close to the flat alveolar fricative class centroid (FAR) and the second close to the nasalised alveolar class centroid (NAN). Several questions arise from this observation. It can be seen that within one phone the output locus is likely to shift according to context, in which case, more than one centroid is required to represent that phone and it is plain that contextual variation must be taken into account when calculating the distance. It also remains to correlate the objective distance metric with perceived quality. These issues are the subject of ongoing work. As a final point, it is interesting to note that a plot of the type shown in figure 7 can give visual feedback that may be of use as a rehabilitation aid in the speech therapy clinic .

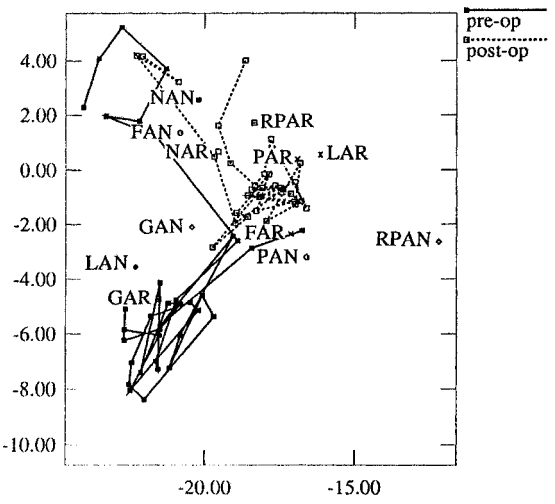


Figure 7. Pre-op and post-op recordings of "s" plotted in the two most discriminant dimensions of the voiceless fricative feature space.

5. SUMMARY

A method for measuring the accuracy and separability of voiceless fricatives has been proposed. It consists of broad identification of voiceless speech frames followed by formant analysis. Formant feature vectors are then transformed into a voiceless fricative feature space which is optimised to discriminate between several acoustically and perceptually distinct sounds which belong to the class. This space is designed to permit distances between post-operative and pre-operative patient speech samples to be measured. Further work is required to take account of legitimate contextual variation within a phone before an objective quality measure can be extracted.

ACKNOWLEDGEMENTS

This project is funded by the British Cancer Research Campaign.

REFERENCES

[1] E. Freedlander, C.A. Espie, L. Campsie, D.S. Soutar, and A.G. Robertson. "Functional implication of major surgery for intra-oral cancers" *British Journal of Plastic Surgery*, 42, pp. 266-269, 1989.

[2] C.A. Espie, E. Freedlander, L. Campsie, D.S. Soutar and A.G. Robertson. "Psychological distress in patients undergoing major surgery for intra-oral cancer." *Journal of Psychomatic Research*, 1989.

[3] A. Gnanadesikan et al. "Discriminant analysis and clustering" *Statistical Science*, Vol. 4, No. 1, pp 34-69, 1989.

[4] A. Crowe and M.A. Jack. "Globally optimising formant tracker using generalised centroids" *Electronics Letters*, Vol. 23, No.19, pp 1019-1020, 1987.

[5] J.E. Shoup, N.J. Lass and D.P. Kuehn. "Acoustics of Speech" *Handbook of Speech-Language Pathology and Audiology*, Ch 6, pp. 171-190, 1988