EVALUATION OF A SYSTEM FOR SEGMENTAL SPEECH QUALITY ASSESSMENT: VOICELESS FRICATIVES

Alan A. Wrench,
CSTR, University of Edinburgh, Edinburgh, UK.
email: aaw@cstr.ed.ac.uk www: http://www.cstr.ed.ac.uk
Mary S. Jackson, David S. Soutar,
Canniesburn Hospital, Bearsden, Glasgow, UK.
A. Gerry Robertson,
Beatson Oncology Centre, Western Infirmary, Glasgow, UK.
Janet MacKenzie Beck
Dept. Speech & Language Sciences, Queen Margaret College, Edinburgh, UK.

ABSTRACT

In this paper, an automatic assessment procedure for monitoring the speech progress of patients who have undergone intra-oral surgery is evaluated. The metric is currently limited to fricative segments and is based on centroid analysis of the speech spectra. The scores provided by this means are correlated with an articulatory assessment provided by a panel of 3 phoneticians. Preliminary results show a degree of correlation between mean panel scores for perceived quality and scores relating to this measure obtained by computer analysis but lateral and nasal co-articulation are not well represented by the metric.

1. INTRODUCTION

It is intended that computer analysis should provide a consistent and objective appraisal of segmental speech quality related to articulatory function and should be used for comparing the speech outcomes of patients who have undergone different treatment modalities. It is particularly useful in its potential to provide consistent appraisal across a number of surgical centres and, most crucially, to reduce to a fraction the number of person hours required to complete an articulatory assessment by other means.

1.1 Patient speech database

The speech database consists of 6 sentences:

The price range is smaller than any of us expected.
They asked if I wanted to come along on the barge trip.
Amongst her friends she was considered beautiful.
John could lend him the latest draft of his work.
[From forty love the score was now deuce and the crowd grew tense.
***or** I think I see the sun shining on some thin fish in the loch.]*
The bulb blew when he switched on the light.

which were recorded in sessions timed as specified in table 1. The speech recordings are digitised at 16 bits per sample at 20kHz with an effective bandwidth of 16kHz. The age range of patients is 29-88 and the ratio of males and females is approximately 2-1 (although 10-2 for the sample used for this paper). At the time of referral the tumour stage T1-T4 is distributed as 1-1-2-2. Tumour site is equally distributed between the three areas: Tongue, floor of mouth and retromolar trygone. Approximately 216 patients have been enrolled at this time.

Session	Timing	Condition
1	Pre-op	At the time of biopsy
2	2-5 weeks post-op	After removal of tracheostomy tube
3	4-8 weeks post-op	Prior to radiotherapy(XRT)
4	15-18 weeks post-op	1-3 weeks post-XRT
5	18-22 weeks post-op	3-8 weeks post-XRT
6^{+}	4-6 week intervals	While speech therapy required

Table 1. Timing Criteria for recordings.

2. PATIENT GROUP

The patient group studied in this paper comprises of 12 patients who had attained 5 or more recording sessions. Table 1. indicates the site and staging of the tumour and the type of surgery and reconstruction.

Code	Stage	Site [1]	Surgery[2] & Reconstruction[3]
AEM	T2N0	Right Tongue & F. of M	M.O., F.N.D.,H.G., Reconstruct F.R.F.F.
CMM	T4N3	Left R.M.T & soft palate	M.E., F.N.D., Reconstruct F.R.F.F.
DJD	T2N0	Right Tongue	F.N.D. & F.R.F.F.
DNR	T2N0	Anterior F.of M.	R.R.A. & Bilateral N.L.F
EXL*	T2N0	Right Tongue	F.N.D.,H.G.,F.R.F.F.
FMW	T4N1	Left Tongue & F. of M.	F.N.D., M.E., & F.R.F.F.
HXA	T3N0	Left Tongue	Quilted S.S.G
JIY	-	palatal	S.S.G. & dental plate
MXM*	T2N0	Right F. of M.	M.E., F.R.F.F.
PXM	T1N0	Anterior tonsillar pillar	Direct Closure
RDM	T2N0	Anterior F. of M. & tongue	Bilateral N.L.F.
RED	T4N0	Left R.M.T.	F.N.D.

Notes: * Female patients
1. F. of M. Floor of Mouth
R.M.T. Retromolar Trygone

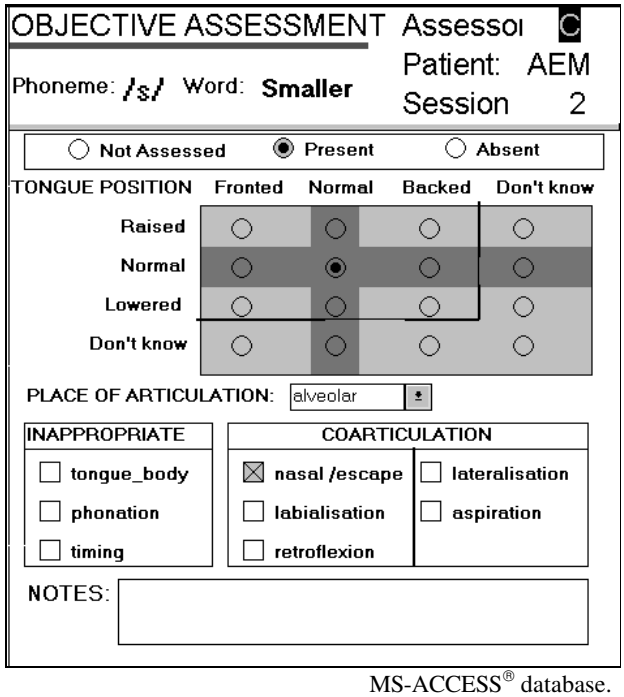
Surgery always consists of excision of the tumour and surrounding tissue and associated procedures noted in each case.

2. F.N.D. Functional Neck Dissection
   M.O. Mandibular Osteotomy
   R.R.A. Rim Resection Alveolus
   H.G. Hemiglossectomy
   M.E. Mandibular Excision
3. S.S.G. Split Skin Graft
   F.R.F.F. Free Radical Forearm Flap
   N.L.F. Nasolabial Flap

Table 2. Pathology and surgical details for each patient

3. PANEL ASSESSMENT

The panel assessment is based on categorising perceived articulation, particularly with reference to perceived lingual articulation.. Both analyses are carried out on 6 /s/ segments for each of 12 patients in 5 progress monitoring sessions. The segments came from the words: pri**c**e, **s**maller, a**S**ked, among**S**t, con**S**idered taken from the set of six sentences.



MS-ACCESS® database.

Figure 1. Articulatory assessment form

The three members of the assessment panel were phoneticians with prior segment labelling experience.

3.1 Subjective Assessment

The subjective score was awarded to each session by the assessor after listening to the sentences and completing the articulatory assessment. The scale was defined as follows:

1. Normal accent
2. Common speech defects e.g. slurring, lisping
3. Can tell there is something not quite right
4. Thoroughly odd sounding
5. Upsetting to listen to

3.2 Articulatory Assessment

The amount of speech data provided by the patients was insufficient for the articulatory test used by Pauloski and Logemann[3][4] and so the following protocol was implemented for the purpose of this study. Articulatory assessment was completed using the input form shown in figure 1. and playback software which enabled fast access and repeated listening to segments. The assessment of 6 fricative segments from 3 sentences across 5 sessions took each panel member approximately 0.5hr per patient. The level of concentration required for the task was high. A numerical score per segment was derived from this assessment according to the following formula.

Artic. = (deviation from normal place of articulation [0/1]
+ inappropriate tongue body position [0/1]
+nasalisation [0/1]
+ labialisation [0/1]
+retroflexion [0/1]
+lateralisation [0/1]
+aspiration [0/1])*[scale factor]
where scale factor=4

The mean score was then calculated per segment across the three assessors and the mean of the resulting 6 segment scores was recorded for each session.

It was found that the assessors never identified more than 3 of the 7 available articulatory deviations. It seems likely that this is because it is difficult to discern multiple coarticulations from the audio recording. Consequently, even for the worst speaker, mean scores never exceeded 5 (after amplifying the scale by 4).

4. AUTOMATIC ASSESSMENT

4.1 The distance metric

The automatic analysis of voiceless fricative segments is based on the determination of the frequency and standard deviation of the centroid of the spectral distribution sampled at 4ms intervals throughout a segment. As shown in Figures 2. and 3, this metric permits separation of several perceptually distinct voiceless fricative realisations. Each frame is 12.8ms long.. The application of a 97% pre-emphasis value produces a tighter clustering of dental and labiodental fricatives near the top of the analysis band. Without pre-emphasis these categories are spread over a wider mid-range of frequencies. Frequencies below 500Hz are excluded from the centroid calculation so that breath noise does not affect the metric.

The mean and covariance matrix are calculated for the cluster of points resulting from the frame-by-frame segment analysis in this 2-D feature space. A mean value and covariance matrix calculated from several pre-operative segments is stored for each patient. The distance between post-operative segments and this pre-operative patient specific target is evaluated using the Mahalanobis distance.

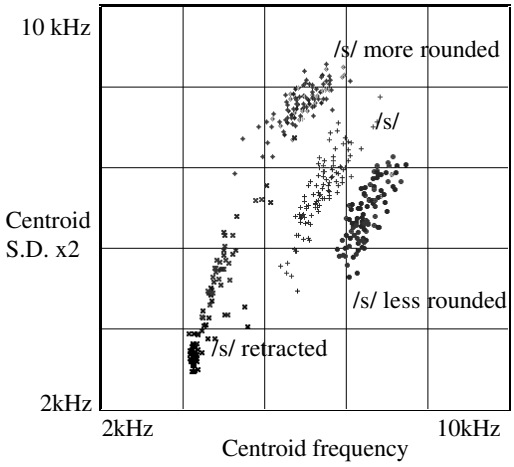


Figure 2 Fricative feature space showing separation of allophones of /s/ by male speaker

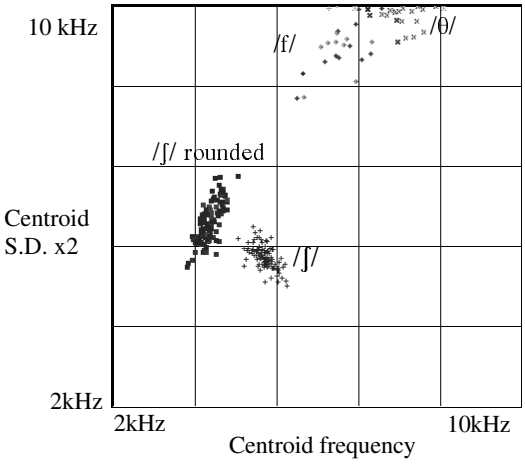


Figure 3. Fricative feature space showing separation of /ʃ/, /f/ and /θ/ by male speaker

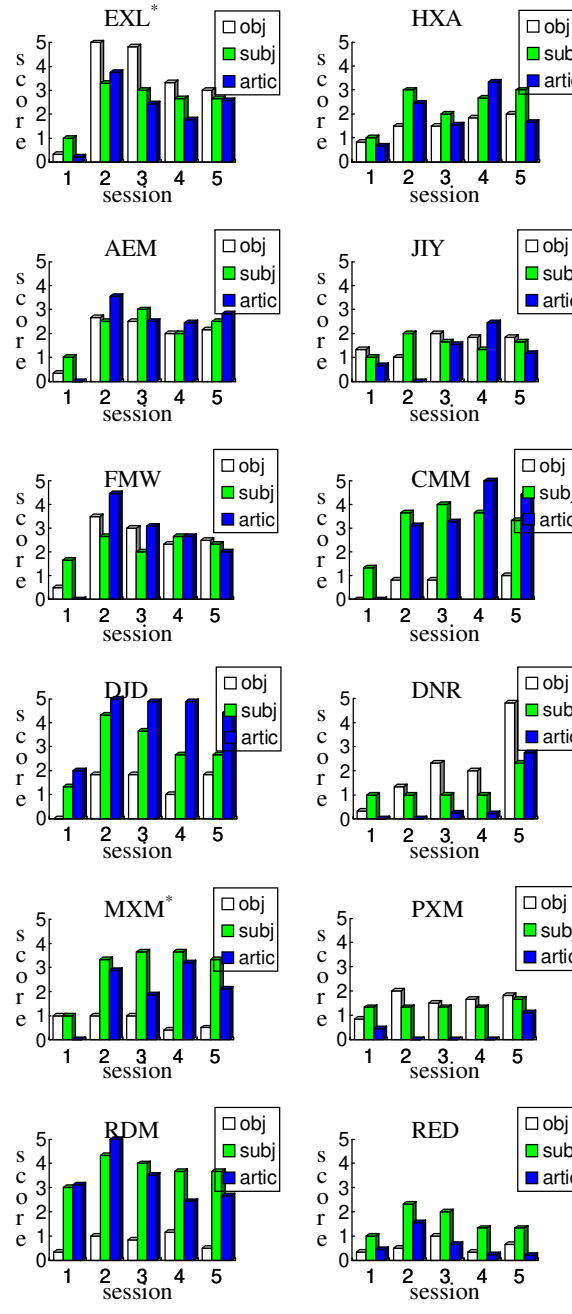
4.2 The assessment procedure

The patient speech data is stored digitally on a PC. The software permits the operator to select any session by any patient and they are then presented with the texts of the sentences stored for that session. Selecting each sentence in turn the operator is presented with the waveform, a spectrogram and a voiced unvoiced decision chart, all calculated in real time. By sweeping the cursor over regions of these plots the operator can select the segment of interest, helped by the facility to playback the highlighted region. The fricative analysis itself is automated to the extent that all voiced frames are rejected. This means that, except in fricative clusters, the segment selection can extend beyond the actual segment boundaries, reducing the precision required by the operator for this task. The analysis procedure is completed at the click of a button and the operator is presented with the score.

The target is determined by analysing six pre-operative segments and storing the accumulated mean and covariance matrix, again simply by clicking a button.

The assessment of 6 /s/ segments from 3 sentences across 5 sessions took the operator approximately 20 mins per patient. The level of concentration and experience required for this objective assessment was considerably less than for the articulatory assessment.

5. RESULTS



Notes: * female
MXM Lateralisation not picked up by objective measure
CMM Velopharyngeal incompetence not picked up by objective measure.
RDM Poor pre-operative speech target

Figure 4. Scores for the Automatic Objective Assessment of 6 /s/ segments compared with panel assessment of the articulation of these segments and an overall subjective score of the speech quality.

Comparative results of the automatic objective assessment and the subjective and articulatory assessment scores are graphed in figure 4. for each session and each patient.

6. DISCUSSION

6.1 Choice of metric

The use of the centroid for fricative identification is an established metric. The second moment appears to give a crude but effective indication of spectral spread. Forrest et al [5] found that the second moment was redundant when a metric was derived from the first 4 moments of the spectral distribution but we have not utilised the higher moments and cannot confirm this finding. Another possible alternative would have been a filterbank representation with 12-14 mel-scaled channels between 500Hz and the upper band limit. currently under investigation. The difficulty with this high dimensional approach is that discriminant training is required to accentuate changes in spectral content which are important for discriminating between categories of fricative and the choice of speaker independent discriminative categories is non trivial. Using speaker-dependent feature spaces is not an option since this places the complex discriminative training process within the assessment procedure. The centroid feature space has the advantage of a continuous space which can be visualised in 2 dimensions. Furthermore, this space seems capable of distinguishing key fricative articulations. However there are 2 important categories on which it fails. Firstly, lateral fricatives span a wide area of the feature space depending on how far back the lateral constriction is located and the degree of lip rounding (as is the case for /s/, Figure 3). To find a method of distinguishing between lateral fricatives and other categories, some means of determining the number (typically 1-2 for /s/ and /2-4/ for lateral fricative) and frequency of formants is desirable. This might be achieved with the use of the filterbank based metric or by multiple centroid analysis [2][6]. The second category consists of nasalised fricatives, which exhibit weaker frication. It may be possible to include normalised energy as an added dimension to the distance metric inorder to distiguish this category.[*]

6.2 Segmental problems exhibited by intra-oral patients

Patient's with node involvement will sometimes incur damage to the hypoglossal nerve during surgery (e.g. CMM) which results in velopharyngeal incompetence and associated poor speech quality resulting from absence of stop closure and weak obstruents. Three patients in this study had good speech quality following surgery: JIY, PXM and RED had no tongue or nerve involvement. DNR had no direct tongue involvement but before session 5 disintegration of the nasolabial flap resulted in a bifid tongue and speech quality deteriorates.

6.3 Evaluation of the objective score

An important property of the automated measure is objectivity. The repeatability of the measure by different operators should provide the same results. This was achieved by using standard texts and an articulatory target calculated across specified segments.

Patient specific targets, while providing a realistic goal for post-operative production, are not ideal. A significant number of patients have pre-operative speech quality which is not acceptable (e.g. RDM, DJD). This can be due to the presence of the tumour or having made the recording after the biopsy. In the case of RDM the objective scores falsely indicate good post-operative speech quality due to the erroneous target. It would be preferable, in such cases, to have a normal target. Indeed it may be preferable to have patient independent targets for all patients.

The measure is also apt to penalise patients with a large articulatory range. One method of redressing this anomaly is to provide a set of segment specific targets. This would only be practicable, however, if patient independent targets were introduced.

7. ACKNOWLEDGEMENTS

This work was funded by the British Cancer Research Campaign.
Assessment panel members: J. Scobbie, M. Nairn, S. Fitt

8. REFERENCES

[1] A.Wrench, J. Laver, M. Jack, G. Robertson, D. Soutar, M. Jackson, J. Beck, *Objective Speech Quality Assessment in Patients with Intra-Oral Cancers: Voiceless Fricatives*. Second International Conference on Speech and Language Processing ICSLP-92, Banff, Vol. 2, pp 1071-1074, October 1992.

[2] A.Wrench, M. Jackson, M. Jack, G. Robertson, D. Soutar, J. Beck, J. Laver. *A Speech Therapy Workstation for the Assessment of Segmental Quality: Voiceless Fricatives.* 3rd European Conference on Speech Comm., Tech., Berlin, September 1993

[3] B.R. Pauloski, J.A. Logemann, et al. *Speech and Swallow Function After Tonsil/Base of Tongue Resection With Primary Closure*. Journal of Speech and Hearing Research, Vol. 36, 918-926, October, 1993.

[4] J.A.Logemann, B.A. Pauloski, et al. *Speech and Swallowing Function After Anterior Tongue and Floor of Mouth Resection With Distal Flap Reconstruction*. Journal of Speech and Hearing Research, Vol. 36, 267-276, April 1993.

[5] K. Forrest, G. Weismer, P. Milenkovic and R.N. Dougall. *Statistical analysis of word-initial voiceless obstruents: Preliminary data*. J. Acoust. Soc. Am., vol. 84, pp. 115-123 1988.

[6] A.A.Wrench. *Analysis of fricatives using multiple centres of gravity*, Proc ICPHS95, Stockholm 1995.

[*] Normalised energy is already incorporated as part of the voiced/unvoiced detection algorithm [1].