SIGNAL PROCESSING FOR USE IN THE ASSESSMENT OF DYSARTHRIC SPEECH

W.A. Simm*[†], P.E. Roberts* and M.J. Joyce*

*The Nuclear and Biomedical Engineering Research Group, Department of Engineering, University of Lancaster, UK
[†]Email: w.simm@lancaster.ac.uk, Tel: +44 (0)1524 5 93325, Fax: +44 (0)1524 381707

Keywords: Speech, Dysarthria,

Abstract

This paper details the signal processing techniques used to produce novel speech metrics for the quantitative assessment of dysarthric speech. A number of different processing methods are used to produce measures designed to aid speech therapists and patients alike in therapy and the assessment of speech quality. The three measures also have the potential for diagnosis of condition and tracking of trends in speech.

1 Introduction

The rationale of this research is that signal processing techniques could give new and innovative speech metrics relevant for speech therapy. Discussions with speech and language therapists (SLTs) have resulted in research into new metrics of speech to help with diagnosis and therapy for speakers with dysarthric speech disorders.

This paper outlines research into investigating the differences in word closure between dysarthric and control speakers. A possible measure as a result of this analysis is proposed. This paper also describes the signal processing algorithms used by two prototype computer-based dynamic speech measures. These measures have been designed to assist clinicians in their assessment and patients in improvement of speech. The two prototypes, a rate/gap display and a cumulative frequency plot, have received favourable reactions from SLT professionals.

The premise of the first measure discussed is that a reduction in control of the speech production organs introduced by dysarthric conditions may be visible in the speech waveforms produced. This will allow tracking of changes in condition, and the possibility of early diagnosis. The initial study has been to specifically investigate the way in which words are closed.

Therapists currently exploit speech rate and the ratio between durations of utterances and silences. The second measure, a rate / gaps display to present this information automatically, is discussed here. Signal processing is required to determine the presence of a fundamental voiced frequency component in the spectrum of the sound. This allows the identification of word boundaries in the speech, allowing a word count and speech:silence ratio to be determined. This information is presented as a dynamic vector. The vector is recalculated at a predefined frequency so that its length, position and variability will give dynamic information to the observer.

A common method of evaluating speech samples is to break the speech down into its constituent formant frequencies. This type of analysis is used traditionally for specific sounds, however here the same technique is used over much longer speech samples. The premise of the third measure discussed is that the physical influences caused by the dysarthric conditions will result in specific effects on the formant frequencies of speech. These effects may introduce differences from those of normal speech. Software has been produced to plot cumulative three dimensional plots of the lowest formant F1 against the next higher formant F2. Differences in plots can be noted in samples produced from people with different conditions. The deterioration in a person's speech quality can be tracked by noting differences in the data plots.

The rate / gaps display and cumulative formant plots were first proposed by Roberts [14], and furthered by Simm *et al* [18]. Application of signal processing techniques has been carried out on Parkinsonian speech by Harel *et al* [8], where changes in the variability of the first formant frequency are proposed as a potential biomarker of early disease progression. Dysphonic symptoms in speech from Multiple Sclerosis patients have been assessed using spectrographic analysis by Feijó *et al* [7]. Max and Mueller [12] have researched the ageing effects on the first formant frequency of speech.

The first section of this paper describes the previously published background to the work [14,18], which includes a description of the three prototype measures. This is followed by a section documenting the proposed signal processing methods for the word closure analysis, and the signal processing methods used by the prototype measures.

2 Background

A speech disorder is defined as a "defect or abnormality that prevents an individual from communicating by means of spoken words" [13]. Disorders are caused by hearing loss, neurological disorders, brain injury, drug abuse, physical impairments, vocal abuse, but often the cause is unknown.

2.1 Dysarthria

Dysarthria is defined as a speech disorder resulting from paralysis, weakness or in-coordination of the speech musculature that is of neurological origin. A number of subsystems make up the speech system, and in order for the speech produced to be clear, they must be coordinated with each other and work together. A motor disturbance in any one of respiratory, resonance, articulation or prosody systems can result in dysarthria. In adults, dysarthria can be caused by stroke, degenerative disease, infections, brain tumours, and toxins, and other neurological conditions [11].

Symptoms of dysarthria include slow, weak, imprecise or uncoordinated speech, generally referred to as "slurred". Dysarthric speakers often appear misunderstood because people listening often pretend they have understood rather than suffer the embarrassment of asking the dysarthric speaker to repeat what they have said.

Dysarthria is one of the most common speech disorders and was selected for study here for this reason. It is also seen as a condition where the fundamental conceptual communication and language faculties of the speaker are intact, but the communication problem is caused by the actual articulation or production of the speech [14]. Given the increasing ageing population, the prevalence of dysarthria in the population is likely to increase.

2.2 Current Therapeutic Practice

A major part of the SLT's role is to continuously assess the condition found in a systematic and structured way. This will mainly involve assessing speech performance, but also includes strongly-related aspects such as breathing, posture, general communication capability and alertness to surroundings. Frameworks appropriate for dysarthria assessment exist which are well respected and referred to by SLTs [6].

Therapy for dysarthric speakers usually involves performing speech exercises with the aim of improving or coordinating the speech subsystems. They are encouraged to take frequent pauses for breath, to over-articulate, or to pause before important words to make them stand out [20]. Exercises designed to strengthen the muscles of the face may also help if there is muscle weakness. Software products, which are designed to aid therapy, automate some of the pacing and prompting techniques that are used by therapists and patients [5].

Speech assessment is often based on the subjective judgement by one or more SLTs. The value of analytical quantitative measures is often acknowledged by the speech therapy community. They are also keen to stress the importance of making any such measure accessible to their level of technical ability.

2.3 Data Collection

A dysarthric speech database has been compiled by researchers at Lancaster University with the help of SLTs from the Morecambe Bay Primary Care Trust. The SLTs are involved in identifying and approaching suitable candidates for collection of speech samples, and act as consultants to the research. Speech samples are collected from dysarthric speakers, based upon the reading of set therapeutic texts. Due to the nature of the conditions involved, the speakers are mostly of elderly age, and tended to tire easily of speaking aloud.

The following procedure and equipment was used to obtain speech samples:

- Speech is recorded using a portable Sony MiniDisc recorder and noise cancellation headset microphone. The MiniDisc format is used for convenience; the ATRAC compression algorithm used by these devices has little or no perceivable effect on the quality of the spoken word.
- Speech is then transferred onto the computer using a Sony MiniDisc deck and optical link to the computer's Creative SoundBlaster Audigy interface.
- Recorded clips are trimmed using the open source Audacity audio editing package [2].

2.4 Word Closure Characteristics

The difference between the way a dysarthric speaker and a control speaker closes a word is marked as can be seen from Figure 1 and Figure 2.

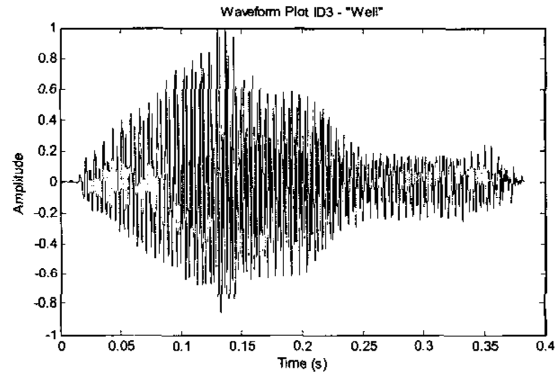

Figure 1: Waveform plot of dysarthric speaker "Well"

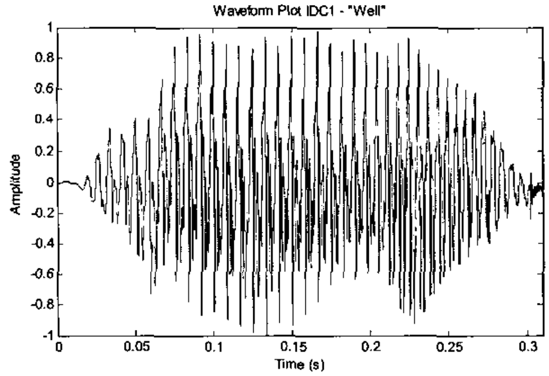

Figure 2: Waveform plot of control speaker - "Well"

The waveform in Figure 2 shows a control speaker speaking the word "well" in the context of the following passage:

"You wish to know all about my grandfather. Well, he is nearly 90 years old. Yet he still thinks as swiftly as ever."

Comparing the control speech in Figure 2 and the moderate-severe dysarthric speech in Figure 1, is can be clearly seen and heard that the dysarthric speaker holds onto the "ell" sound in the word much longer than the control speaker.

It is proposed here that the difference observed in this aspect is a result of reduced control or an in-coordination in control of the glottis organ. This leads to the speaker not being able to close the glottis and hence end the sound effectively.

The potential of this measure is that by detecting the change in a speaker's word closure over time, early diagnosis of dysarthria may be possible and therapy could start before a perceptible audible difference is observed. Since control of the glottis is linked with swallowing difficulties (dysphagia), early diagnosis and therapy could make a large difference to a person's later quality of life.

2.5 Rate / Gaps Display

The premise to this measure was that dysarthria has particular effects on the prosody of the speech [14]. Prosody is the study of characteristics of speech beyond the basic words and spectra, e.g. stress, intonation, tempo.

Therapists currently make use of rate observations, although their measurement is often done by a manual timing of passages of speech [6,15]. The measurement of the ongoing ratio between durations of utterances and silences is also valued by the SLTs, but is difficult to measure without automatic means. Specialist tools are now available which display rate [10] but their use by SLTs appears limited because of cost and portability limitations. A combined display of the rate and gaps parameters has not been found in existing products or discussions [14].

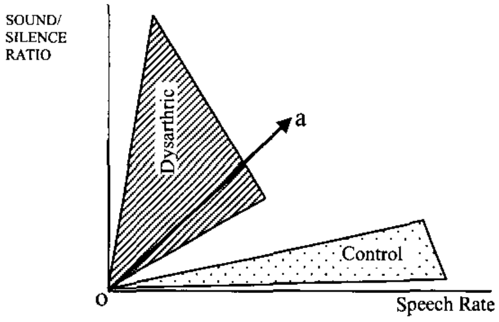

Figure 3: Proposed rate/gaps display [14]

Figure 3 shows an example of this measure. A dynamic vector is implemented, with the horizontal axis corresponding to the speech rate parameter and vertical axis to the ratio of speech:silence (to represent a measure of gaps). The vector, line *o-a* on Figure 3, is recalculated at a predefined frequency so that its length, position and variability gives dynamic information to the observer [14].

To demonstrate the use of this measure, consider a long and stable line running close to the vertical. This would indicate a small proportion of gaps and little change in a slow speech rate. This is shown in Figure 3 as the lined region, and is typical of dysarthric speech. Conversely a long or varying length line running close to the horizontal would indicate a significant proportion of gaps with a rate that was high or varying, respectively. This is shown in Figure 3 as the dotted region, and is typical of control speech

2.5.1 Results

The following was observed [18]:

- All the control speakers produced vectors which kept near horizontal, with the variability being primarily in rate between individuals depicted by variation in length.
- Most dysarthric speakers, showed a higher degree of scatter but with reduced rate range, the vector angle will change, but with constant length.
- Two dysarthric speakers showed extremes of vector position. Further investigation is necessary because the speech quality was judged as very poor. The number of sounds counted (31 and 120) suggested that this may be causing the boundary algorithms to give false data.

Where speech data are available over a period of months, and there was a detectable change in speech characteristics, the trend is also visible on the vector data [14].

2.6 3-D Cumulative formants scatter plots

Evaluation of speech samples is often undertaken initially by breaking the speech down into its constituent formant frequencies. The formant frequencies of speech are found by the spectral analysis of speech signals, which enables the identification of formant frequencies in the speech. The main peaks are referred to as formant frequencies, and are created by the various resonating cavities and wall characteristics in the vocal tract [19]. The premise of this measure is that the physical influences caused by the dysarthric conditions will have specific effects on the formant frequencies that introduce differences from those of normal speech [18].

A common method of interpreting the formant frequency data is to plot the amplitude of the lowest frequency formant F1 against that of the next higher formant F2. This is used traditionally for specific sounds, e.g. vowels; the same technique can be used over much longer speech samples. This allows the comparison of data collected from people with dysarthria caused by different conditions, and the identification of trends within the data [18].

2.6.1 Results

Plotted in this way it is clear to see the three categories of patterns produced as suggested by Roberts [14]. Figure 4.1 shows an example of a reasonably well scattered data set, with no significant clusters. The control samples were consistent with this, along with patients with mild dysarthria. Figure 4.2 shows a uni-polar distribution, indicative of a patient with moderate dysarthria. Figure 4.3 shows a bipolar distribution with the points emphasised into two distinct poles. Patients with this distribution included those with moderate to severe dysarthria [18].

3 Signal Processing

This section describes the proposed signal processing methods for the word closure analysis, and the signal processing methods involved in producing the other metrics described.

3.1 Word Closure Characteristics

The proposed technique to be used to automatically determine the shape of the closure is as follows:

- Filter the waveform to remove low amplitude components of the signal.
- Differentiate the data with the intention of extracting the peaks.
- Correlate the y values to form a plot of the boundary of the waveform.
- Fit a curve to the boundary data and compare with multiple data sets.

3.2 Rate / Gaps Display

The key element in achieving ongoing, real-time measurement of speech rate and the speech/silence ratio of a passage of speech is the identification and timing of the voiced-word sections of sound. Once this has been done, it is possible to establish the word start and end boundaries and hence calculate the rate and ratio information [14].

Three options for this analysis were evaluated by Roberts, who noted that previous research has demonstrated the difficulty of accurate identification of word boundaries [3,14]:

- The absolute sound level determined from the digitised speech data, a threshold level was set by the mean (RMS) level of a period of speech.
- The intensity in dB calculated from the digitised sound levels across a specific window, with the threshold set by the mean of the intensity over a significant period.
- Voiced / unvoiced pitch processing determined from spectral analysis of the speech data.

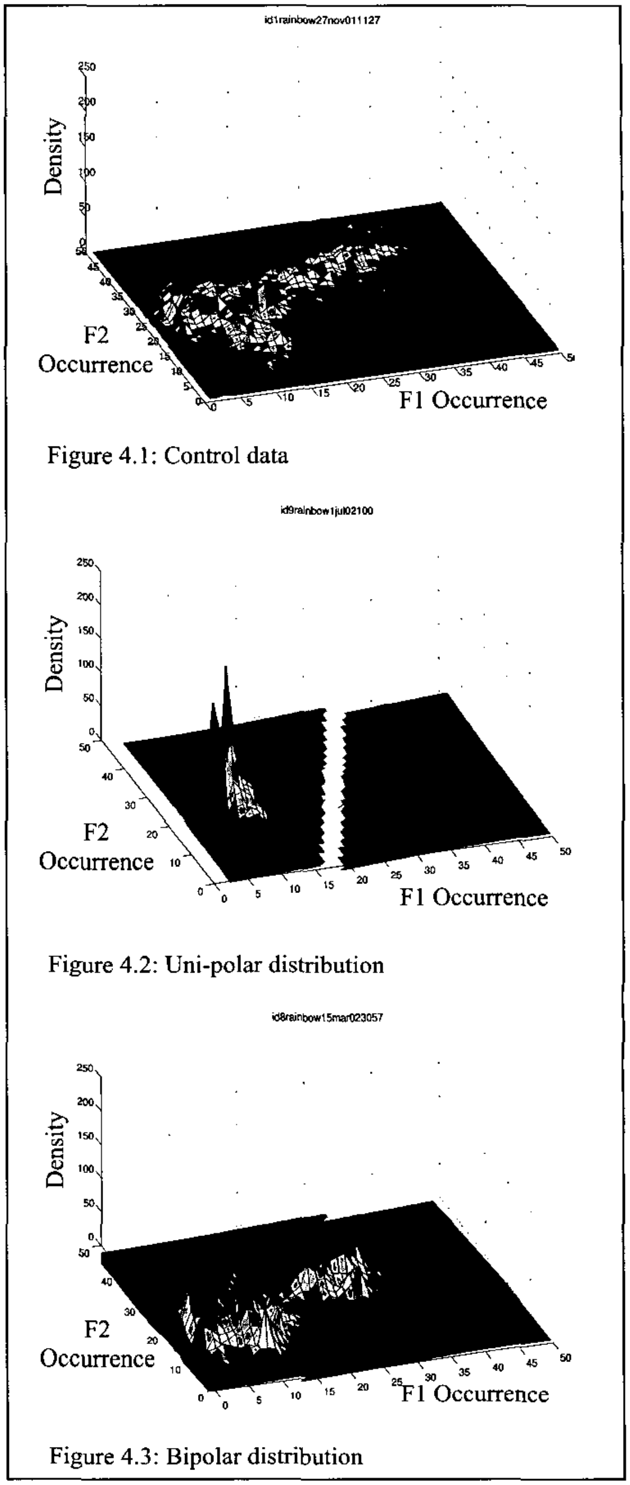

Figure 4.1: Control data

Figure 4.2: Uni-polar distribution

Figure 4.3: Bipolar distribution

Figure 4: 3-D Cumulative Formant Plots [18]

Roberts determined that the voiced / unvoiced processing gave the optimum measure for the vector display [14].

The voiced / unvoiced pitch processing method determines the presence of speech by detecting the presence of a fundamental voiced frequency component in the frequency spectrum of the sound. Spectral analysis of the speech signal enables the identification of frequency peak components [19] and various algorithms have been developed to estimate the pitch [16,9]. An autocorrelation-based algorithm [3] has been used in this work which selects and tracks candidate frequency peaks for the pitch of the voice. The resulting ongoing output from the pitch algorithm is used to decide if the signal is voiced (used in this context to mean a sound being spoken) or unvoiced (used in this context to mean the silence between sounds). This is then used in the subsequent processing to determine word boundaries, rates and ratios and to plot the data [14].

A limitation of using this algorithm, to determine the status of the signal, is that there are some sounds produced which are determined by the absence of pitch to be unvoiced, for example "fin". Also some words have multiple voiced components. However since the algorithm produced the best results in comparative evaluation, it is used here.

A further algorithm is used to determine if the situation has persisted for a period of 100ms or more, if so the status of the signal is determined to be voiced or unvoiced accordingly. This window gives a degree of tolerance to noise, short-term effects or sound drop-outs. A word and silence count can then be maintained, and an average rate ratio periodically calculated for display is then produced [14].

3.3 3-D Cumulative Formant Plots

Estimating the formant frequencies of a speech sample is achieved here by utilising the well-established Linear Predictive Coding (LPC) method with a maximum entropy method (MEM) originally by Burg [1]. The specific implementation used here is from within the Praat phonetic toolset [4].

The LPC method has become increasingly popular for analysing formant frequencies. The resonance frequencies of the speaker's vocal tract are estimated by linear prediction at regular intervals through the signal, and the formants are identified from the peaks [4].

The formant estimating parameters used by the LPC method are documented in Table 1. The pre-emphasis is used to give a more even power distribution across the full speech spectrum. The max formant parameter is used to limit processing requirements. The window parameter sets the duration over which spectrum calculations are made, recognising typical variation of the speech parameters under consideration. The time step parameter determines the resolution of the analysis. Here it is set automatically by the algorithm.

Pre-Emphasis	50Hz
Max Formants	5
Window	0.025s
Time Step	auto

Table 1: Formant Estimating Parameters [14]

Traditionally a scatter plot may be used to plot the lowest frequency formant F1 against the formant with the next highest frequency, F2. However with the analysis being made over much longer speech samples, a different approach is needed to present the data clearly, interpretation of a scatter plot showing so much data is dependent on the resolution of the medium on which it is viewed. The "densities" of the points from a scatter plot are translated to different heights and colours on a 3-d plot, in the hope that more detail in the 3-d shape of the data will be identified [18].

A script was identified [17] and modified to produce 3-d histograms in Matlab. This splits the data into a number of "bins", and maintains a count of how often data fall into a bin. This count is then translated into the height of the 3-D plot. In this case it was determined that 50 bins provided a reasonable resolution. The script was modified to automatically produce graphs of the optimum resolution and scale for the data.

4 Conclusions

Modern signal processing techniques have enabled the implementation here of novel metrics considered to be valuable by SLTs.

The rate / gaps display with the angular movement and varying length of the vector gives near real-time feedback to the observer which may help them pace their speech delivery [14]. Initial consultations with SLTs have been positive; with the impression the tool would be useful. The 3D cumulative formant plots show successfully changes in dysarthria over time and may also aid diagnosis of condition [18]. The metrics outlined are to be developed into a tool designed for use and interpretation by a SLT, however further automatic interpretation of the 3-d plot data would need to be implemented before it would be of use to a SLT. Current research into automating the analysis of the word closure characteristics of speech may lead to further insight into the assessment of dysarthric speech and may have implications in diagnosis of other disorders such as dysphagia.

Acknowledgements

Thanks go to Claire Philpott of the Morecambe Bay NHS Primary Care Trust for her assistance in selecting subjects for speech data collection and assisting collection sessions. This research is funded by the EPSRC.

5 References

[1] N. Anderson. “On the calculation of filter coefficients for maximum entropy spectral analysis”, *Modern Spectrum Analysis*, IEEE Press, pp. 252-255. (1978).

[2] Audacity. “Audacity – The Free, Cross-Platform Sound Editor”, *http://audacity.sourceforge.net/*, (2005).

[3] P. Boersma. “Accurate short term analysis of fundamental frequency and the harmonics-to-noise ratio of sampled data”, *Proc Inst of phonetic science*, **17**, pp. 97-110, University of Amsterdam, (1993).

[4] P. Boersma, D. Weenink. “PRAAT: doing Phonetics by Computer”, *http://www.praat.org,* (2004).

[5] Bungalow Software. “Speech Pacer & Speech Prism”, *http://www.bungalowsoftware.com,* (2003).

[6] P.M. Enderby. “Frenchay Dysarthria Assessment”, *Pro ed,* (1983).

[7] A.V. Feijó, M.A. Parente, M. Behlau, S. Haussen, M.C.D. Veccino, B.C.F. Martignago. “Acoustic Analysis of Voice in Multiple Sclerosis Patients”, *Journal of Voice*, **18**, pp. 341-347, (2004).

[8] B.T. Harel, M.S. Cannizzaro, H. Cohen, N. Reilly, P.J. Snyder. “Acoustic characteristics of Parkinsonian speech: a potential biomarker of early disease progression and treatment”, *Journal of Neurolinguistics*, **17**, pp. 439-453, (2004).

[9] M. Hosseinpour, H. Amindreza. “Pitch Estimation using mMusic algorithm based on the sinusoidal speech model”, *Iranian Telecommunication Research Centre, Iran, http://www.cic.aku.ac.ir*, (1996).

[10] Kay Elemetrics. “Kay Elemetrics Corporation Website”, *http://www.kayelemetrics.com/*, (2001).

[11] R. Love, W. Webb. “Neurology for the Speech-Language Pathologist (2nd ed.)”. *Butterworth-Heinmann*, (1992).

[12] L. Max, P.B. Mueller. “Speaking F_0 and Cepstral Periodicity Analysis of Conversational Speech in a 105-Year-old Woman: Variability of Aging Effects”, *Journal of Voice,* **10**, pp. 245-251, (1996).

[13] NIDCD. “National Institute on Deafness and other Communication Disorders Glossary”, *http://www.nidcd.nih.gov/health/glossary/glossary.asp*, (2004).

[14] P.E. Roberts. “Improvement of Computer Recognition and Development of New Metrics”, PhD thesis submitted Dec 2004. *Lancaster University,* (2004).

[15] S.J. Robertson, F. Thomson. “Working With Dysarthric Clients: A Practical Guide to Therapy for Dysarthria”, *Winslow Press,* (1986).

[16] S. Saito, Nakata. “Fundamentals of Speech Signal Processing”, (1985).

[17] K. Sidaros. “DSP Course 02457 - Hist2d.m”, *http://isp.imm.dtu.dk/teaching/04364/ex2/hist2d.m*, (1999).

[18] W.A. Simm, P.E. Roberts, M.J.Joyce, C. Philpott. “The Quantittive Assesment of Dysarthric Speech for Use in Evidence-Based Speech Therapy”, *HCI International, Las Vegas, USA.* MIRA Digital Publishing CD-ROM, (2005).

[19] K. N. Stevens. “Acoustic Phonetics”, MIT Press, (2000).

[20] L. Styba. “The Speech-Language Pathology Website”, *http://home.ica.net/fred/public_html/anch10-1.htm*, (1999).