THE INTELLIGIBILITY OF SPEECH IN CEREBRAL PALSY: THE EFFECTS OF MANIPULATING THE ACOUSTIC SPEECH SIGNAL

C. McKilligan[1], J. van Doorn[2] & S. Pitt[3]

[1] School of Electrical Engineering
University of Technology, Sydney

[2] School of Communication Disorders
The University of Sydney

[3] The Spastic Centre of New South Wales

ABSTRACT - Lack of speech intelligibility is a problem for many people who have cerebral palsy. This pilot study investigated whether improved speech intelligibility could be achieved by modifying the frequency and durational characteristics of the acoustic speech signal. Formant frequency modifications using linear predictive analysis and synthesis and time scale modifications were used to manipulate the speech signal for token utterances from a single speaker. The resultant acoustic signals were used as the basis for some preliminary listening tests to establish whether improvements in intelligibility occurred. Varying degrees of improvemed intelligibility were achieved, depending on the nature of the modifications. These encouraging results provide justification for an expanded project to investigate optimal speech processing methods using more speakers and a larger corpus of utterances.

INTRODUCTION

The motor disorders which characterise cerebral palsy can affect the function of the muscles involved in the production of speech just as they affect the postural muscles. A significant proportion of the cerebral palsied population have barely intelligible or unintelligible speech because of disorders in the articulatory (lips, tongue and jaw) muscles. Efforts to provide effective means of communication for these people include the development of augmentative communication devices, and more recently speech recognition systems have been investigated for use with the disordered speech of cerebral palsy. Fright, Bates, Briesman, Garden, Kennedy, McCallum, Sewell, Tan & Turner (1985) reported successful implementation of a speech translation device which recognised a word, looked it up in a dictionary and then synthesised a correct version of the word. Sy and Deller (1989) have done similar research into AI generation of messages for speech disabled people. Commercial speech recognition systems are currently being investigated for their suitability with disordered speech (Lariviere, McKinnon & Risebrough, 1993; Rosengren, Raghavendra & Hunnicutt, 1994).

However, the approach in our research was not based on recognition systems but rather on manipulation of the original acoustic signal to improve its intelligibility, an approach similar to that adopted for deaf speech by Maassen & Povel (1985). The correction factors which were investigated for their effect on the speech intelligibility were based on the findings of an earlier study into the acoustic characteristics of cerebral palsied speech (van Doorn, 1983; van Doorn, O'Dwyer & Neilson, 1986). Results from that study indicated that when cerebral palsied speech was compared with fluent speech, it was characterised by a slower speech rate, with longer syllable durations, and formant trajectory patterns with reduced ranges of frequencies and transitions with slower rates of change. Two other findings from that study influenced the approach to the present pilot study. First, the longer syllable durations in cerebral palsied speech maintained relative duration lengths similar to those found in fluent speech. Secondly, the formant trajectory patterns for each cerebral palsied speaker showed a considerable degree of reproducibility for repetitions of the same sentence.

The reproducibility of the formant patterns was critical to the feasibility of this study. Meaningful acoustic changes could only be made if there was consistency in the original acoustic patterns. Reproducibility of EMG recordings in the 1982 study was substantiated by

Neilson & O'Dwyer (1984) and was also reflected in the formant trajectory patterns. Additionally the maintenance of relative syllable durations in cerebral palsied speech meant that uniform time compression could be trialled in the first instance, rather than selective time warping. If uniform time compression coupled with formant frequency expansion were applied to the acoustic signal, the formant trajectory patterns would more closely resemble fluent speech. Subsequently, the rate of formant transition change would also increase. Intelligibility has been linked to the rate of change of the second formant in dysarthric speech (Kent, Kent, Weismer, Sufit, Brooks & Rosenbek, 1989), which provides some support for the notion that time compression and frequency expansion may improve intelligibility.

If speech processing of this nature does in fact improve intelligibility in cerebral palsied speech it leads to the prospect of the development of a "speech translator" which takes barely intelligible speech and converts it to intelligible speech. This speech would not be artificial speech, but rather an altered and intelligible version of the speaker's original utterance. This preliminary study investigated whether the processes of time and frequency warping had a positive impact on the intelligibility of a speaker with barely intelligible speech.

METHOD

The strategy employed in this pilot study was to record the speech of a single subject who met the same perceptual speech characteristics as those specified in the original study, estimate the frequency and time corrections required, and implement the corrections on two test utterances which were then used for preliminary intelligibility testing.

Subject

The subject was an adult male with cerebral palsy who by chance had participated in the earlier (1982) study. One of the investigators (a qualified speech pathologist) judged his speech to be moderately dysarthric with limited intelligibility. Additionally he had no conspicuous respiratory difficulty which interfered with his ability to phonate.

Speech samples

Three different speech samples were recorded, using a Marantz CP430 tape recorder with a Beyer Dynamic M88(N) microphone and TDK metal tapes. The first utterance, "Do all the old rogues abjure weird ladies?" was the sentence used in the original study, which was required for a comparison with the fluent data from that study. Five repetitions of the utterance were recorded. Single utterances of two additional sentences "They will make many friends." and "Keep your house very clean" were also recorded. These were chosen randomly from the Assessment of the Intelligibility of Dysarthric Speech (Yorkstow & Beulkelman, 1981) for use in the preliminary intelligibility tests on the modified speech.

Estimation of correction factors.

Two specific corrections on the intelligibility of dysarthric speech in cerebral palsy were examined in this pilot study: time compression and formant frequency expansion, either individually or together. The correction factors were determined by comparing overall utterance lengths, syllable lengths, and general ratios of formant frequencies for the utterance "Do all the old rogues abjure weird ladies?". A comparison with fluent speech for the same utterance yielded an average time compression factor of 0.667 and an average frequency multiplier for all three formants of 1.16. The corrections were performed as blanket changes to the speech.

Processing strategies

An important consideration in determining the processing strategies to improve intelligibility of the speech of people with cerebral palsy was that the improved speech should be a product of the original voice as far as possible. In this respect, the processing used was steered toward

methods which are capable of analysing and resynthesising speech while retaining the speaker characteristics of the original speech in the resynthesised speech, altering only features affecting intelligibility. Processing strategies were divided into time warping, frequency warping and hybrid processes which combined both time and frequency warping. Frequency warping involved only one strategy, whereas time compression of the speech signal was achieved using three different strategies.

Frequency warping

Correction of formant frequencies was achieved by using Linear Predictive Coding (LPC) techniques (Markel & Gray, 1976). LPC is a particularly suitable form of speech processing for this application because it is possible to analyse speech, alter the formant frequencies and resynthesise, yet still maintain the voice characteristics of the original speech by using residual excitation in resynthesis. The first three formant frequencies were detected, multiplied by 1.16 and then used for resynthesis using the ASL LPC Parameter Manipulation Software, an LPC analysis synthesis program which operates with the Kay DSP 5500 Sonagraph. During analysis the autocorrelation method of analysis with rectangular window weighting was used. Pre-emphasis of 0.95 was applied to voiced periods of speech and 0.6 to unvoiced periods of speech. Frame length for unvoiced frames was set to 20 ms. The filter order was set to 12 poles. Speech was then resynthesised from the modified formants using residual excitation to maintain the original voice characteristics.

Time Warping

Synchronised Overlap-Add Algorithm

Synchronised Overlap Add Algorithms (SOLA) are an efficient method of Time Scale Modification which was developed by Roucos and Wilgus (1985). They are capable of high quality time compression of between 50% and 100% or super compression (< 50%). SOLA algorithms achieve time compression by generating composite speech segments which replace one or more segments to give a time scale reduction. The advantage of SOLA algorithms in this application is its ability to compress speech without altering voice pitch. The main disadvantage is that arbitrary local compression factors can occur because of the synchronisation mechanism. A SOLA algorithm was implemented with a compression factor of 0.667, using a 72ms segment size (which was judged to produce optimum speech quality in this application).

Midvowel compression

The speech utterances were LPC analysed and the vowel regions were identified by examination of the formants and listening to selected areas of the waveforms. Once all the vowel regions were identified the frame numbers in which they occurred were noted, and every second frame in these regions was removed. A higher compression rate was used here to compensate for selective compression of vowels only. This led to the entire utterance compression being close to 0.667. Once frames were removed in the vowel regions identified, the utterance was resynthesised using residual excitation.

Global speech rate increase

An increase in the speech rate was achieved using digital sampling, increasing the sampling rate by a factor of 1.5 (corresponding to the compression factor of 0.667 used in the SOLA compression), and resynthesising at the new sampling rate. The serious disadvantage of this form of time compression was that it also increased pitch and formant frequencies and bandwidths by the same factor, thus negating the criterion of maintaining original voice characteristics as far as possible.

Hybrid processing

Global speech rate increase with LPC frequency restoration.

The sampling rate was increased by 1.5, causing formant frequency and bandwidths to increase by the same amount. The formant frequencies were reduced to the desired 1.16 factor. F0 and the formant bandwidths were scaled by 0.667 to reduce them to their original values before the sampling rate increase. The utterance was resynthesised using residual excitation.

LPC Formant frequency Adjustment and SOLA

This consisted of formant adjustments followed by SOLA algorithm. Processing was completed in this order to ensure that discontinuities introduced by the SOLA processing did not affect the LPC analysis and synthesis, as SOLA is the least exact of the two algorithms.

Preliminary Intelligibility Testing.

Processing strategies as previous established were grouped as follows: No processing, Time warping, frequency warping, and hybrid warping. Six listening tapes were prepared, with one sample sentence selected randomly from each of the process groups for each tape. The order of the process groups was randomised also. Because there were only two different utterances, a measure to reduce learning effects was incorporated by adding three fluent utterances (also from the AIDS instrument) to alternate with the cerebral palsied speech. Thus there was a total of seven sentences per tape. Each sentence was recorded on the tape twice in succession with silence of 5 seconds between every utterance on the tape. Ten listeners were used to listen to each tape and record what they heard. Listening tests were scored by marking syllables correct. The % syllables correct was scored as a total score across all listeners for each process of each of the two sentences.

RESULTS.

Results for the two sentences "They will make many friends" and "Keep you house very clean" are presented in Table 1. Improvements in intelligibility showed similar patterns for both sentences, with substantially increased "%syllables correct" scores for three processes: The hybrid LPC formant correction and SOLA time compression, midvowel compression, and speech rate increase, which showed most improvement.

Process	Type	% syllables correct	
		Sentence 1	Sentence 2
Unmodified	Unmodified	33.3%	9.6%
SOLA compression	Time compression	36.7%	1.5%
LPC formant correction	Formant modification	33.8%	9.4%
LPC formant correction and SOLA compression	Hybrid	45.0%	19.0%
Midvowel compression	Time compression	41.7%	30.0%
Speech rate increase	Time compression	54.5%	51.5%
Speech rate increase with formant frequency restoration	Hybrid	6.8%	4.2%

Table 1. Results of intelligibility tests for various forms of processing. Sentence 1 is "They will have many friends" and Sentence 2 is "Keep your house very clean."

DISCUSSION

Cautious interpretation of the results revealed that changing the acoustic characteristics of dysarthric speech for a single speaker resulted in improved intelligibility, for at least three of the processes which were trialled. The preliminary nature of the investigation meant that there were limitations in the study in both the implementation of the speech processing and the interpretation of the intelligibility tests. Several difficulties were encountered during the preparation of the speech samples for intelligibility testing. The Kay Sonagraph and ASL Software restricted the amount of speech which could be processed to about 2.5 seconds, so that the sentences had to be processed in two sections. Additionally the formant tracking algorithms in the ASL software often marked F2 as F1, and thence F3 as F2. This was compensated by using the same multiplicative factor for all formants, so that further errors were not introduced in the resynthesis. Despite trialing many different analysis and resynthesis options which were available on the ASL software, it was difficult to obtain high quality synthesised output - an effect which was compounded when alterations to formant structures were performed and very evident when processing CP speech.

During the investigation it was clear that the intelligibility scores were influenced by the poor and variable quality of the resynthesised speech as much as the acoustic changes which were made. For instance, it was found that speech rate increase (a process which was minimally affected by the effects of speech processing) yielded the best improvement in intelligibility. However, when formant and pitch corrections were applied to compensate for concomitant increases in pitch and formant frequencies, the intelligibility scores dropped dramatically, probably as a result of poor quality synthesis. It is anticipated that hybrid formant correction and SOLA compression would be capable of yielding similar results to the speech rate increase provided that the quality of the LPC synthesis were improved. Clean processing of the signal that does not introduce side effects will be vital in any continuing investigation. It is envisaged that, were the results of LPC synthesis improved to provide a cleaner resynthesis, much improved intelligibility scores would occur in several of the processes.

The small number of different sentences available for the intelligibility tests meant that the results were confounded by the possibility of learning effects of the listeners. However, the randomisation of the order of the differently processed sentences across six different listening tapes, and the interspersion of additional fluent sentences on the tape minimised any learning effects. Nonetheless, any interpretation of these results must acknowledge the possible learning influences. This limitation can be easily addressed in future studies by the incorporation of a larger corpus of speech utterances.

In summary, the results from this pilot study provide sufficient justification to further investigate the possibility of using acoustic manipulations to improve intelligibility of dysarthric speech. In particular, optimal speech processing with high quality synthesis needs to be developed and tested, using more speakers and a larger corpus of utterances.

REFERENCES

Fright, W.R., Bates, R.H.T., Briesman, N.G., Garden, K.L., Kennedy, W.K., McCallum, B.C., Sewell, M.R., Tan, K.P. & Turner, S.G. (1985) *Computer translation of impaired speech*, TADSEM 85: Australian Seminar on Devices for Expressive Communications and Environmental Control, Sydney.

Hardam, E. (1990) *High quality time scale modification of speech signals using fast Synchronized-Overlap-Add Algorithms*, International Conference on Acoustics, Speech and Signal Processing, Albuquerque, NM.

Kent, R.D., Kent, J.F., Weismer, G., Sufit, R.L., Brooks, B.R., and Rosenbek, J.C. (1989) *Relationships between speech intelligibility and the slope of second formant transitions in dysarthric subjects*, Clinical Linguistics and Phonetics, 4, 347-358.

Lariviere, J., McKinnon, E. & Risebrough, N. (1993) *Is speech recognition worth it?* Speech and Language Technology for Disabled Persons, Proceedings of a European Speech Communication Association Workshop, Stockholm, 95-98.

Maassen, B. & Povel, D.J. (1985) *The effect of segmental and suprasegmental corrections on the intelligibility of deaf speech*, Journal of the Acoustic Society of America, 78, 877-886.

Markel, J.D. & Gray, A.H. (1976) *Linear prediction of speech,* (Springer Verlag: Berlin).

Neilson, P.D. & O'Dwyer, N.J. (1984) *Reproducibility and variability of speech muscle activity in athetoid dysarthria of cerebral palsy*, Journal of Speech and Hearing Research, 27, 502-517.

Rosengren, E., Raghavendra, P. & Hunnicutt, S. (1994) *How does automatic speech recognition handle dysarthric speech?* FONETIK 94, Papers from the 8th Swedish Phonetics Conference, Lund, Sweden, 112-115.

Roucos, S. & Wilgus, A.M. (1985) *High quality time-scale modification for speech,* IEEE International Conference on Acoustics, Speech and Signal Processing, Tampa, Florida, 493-496.

Sy, B.K. & Deller, J.R. (1989) *An AI-based communication system for motor and speech disabled persons: Design methodology and prototype testing,* IEEE Transactions on Biomedical Engineering, 36, 565-571.

van Doorn, J.L. (1983) *Speech Waveforms in cerebral palsy - An acoustic analysis,* Journal of the Acoustic Society of Australia, 11, 17-23.

van Doorn, J.L., Neilson, P.D. & O'Dwyer N.J. (1986) *Dysarthric speech in cerebral palsy - A hornet's nest of acoustic and electromyographic data,* Proceedings of the First Australian Conference on Speech Science and Technology, Canberra.