Effects of Acoustic Features Modifications on the Perception of Dysarthric Speech - Preliminary Study (Pitch, Intensity and Duration Modifications)

T B Ijitona*, J J Soraghan*, A Lowit†, G Di-Caterina*, H Yue*

**Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, United Kingdom, †Department of Speech and Language Therapy, University of Strathclyde, Glasgow, United Kingdom, tolulope.ijitona@strath.ac.uk*

Keywords: Stress marking, perception, dysarthria, acoustics

Abstract

Marking stress is important in conveying meaning and drawing listener's attention to specific parts of a message. Extensive research has shown that healthy speakers mark stress using three main acoustic cues; pitch, intensity, and duration. The relationship between acoustic and perception cues is vital in the development of a computer-based tool that aids the therapists in providing effective treatment to people with Dysarthria. It is, therefore, important to investigate the acoustic cues deficiency in dysarthric speech and the potential compensatory techniques needed for effective treatment. In this paper, the relationship between acoustic and perceptive cues in dysarthric speech are investigated. This is achieved by modifying stress marked sentences from 10 speakers with Ataxic dysarthria. Each speaker produced 30 sentences using the 10 Subject-Verb-Object-Adverbial (SVOA) structured sentences across three stress conditions. These stress conditions are stress on the initial (S), medial (O) and final (A) target words respectively. To effectively measure the deficiencies in Dysarthria speech, the acoustic features (pitch, intensity, and duration) are modified incrementally. The paper presents the techniques involved in the modification of these acoustic features. The effects of these modifications are analysed based on steps of 25% increments in pitch, intensity and duration. For robustness and validation, 50 untrained listeners participated in the listening experiment. The results and the relationship between acoustic modifications (what is measured) and perception (what is heard) in Dysarthric speech are discussed.

1 Introduction

Dysarthria is a neurological disorder that affects the production of sound due to the weakness of the muscles and nerves involved in speech production [1]. The effects of dysarthria are noticeable in speed, pitch variability, consistency and control in speech production. Dysarthria is classified into six distinct types: hypokinetic, hyperkinetic, ataxic, flaccid, spastic and mixed dysarthria. Dysarthria is commonly accompanied with facial drooping, slow speech rate, voice quality deficiencies, lopsided rhythm, slurred speech, increased pitch variability and low loudness [2].

Ataxic dysarthria, as one of the types of dysarthria, is caused by lesions to the cerebellar or the control circuit of the cerebellum affecting its functioning [1]. The causes of dysarthria range from cerebellar damage from stroke, hypothyroidism, cerebral palsy, multiple sclerosis, postoperative trauma to toxicity [1]. Research [3] has shown that the speech subsystems predominantly affected by ataxic dysarthria are phonation, prosody, and articulation [3].

However, research [4] has also shown that prosody plays an important role in the conveying contrastive information, distinguishing between statements and questions, and showing emotions, expressions and attitudinal state of mind. Prosody aids the comprehension of listeners by marking the boundaries, emphasis, and stress thereby reducing ambiguity [4]. Acoustically, prosody is measured by extracting the intensity, fundamental frequency, and duration of speech signals. Over the past few decades, researchers have studied prosody in ataxic dysarthria based on the perceptual and acoustic analysis. One of the motivations for this study is the inaccuracy in prosodic deficiency measurement in ataxic dysarthria such as pitch breaks, pitch variability, equal or excess stress, prolonged syllables, mono loudness, mono pitch and speech slow rate[3, 5].

Recent research work has offered more light to the understanding of prosody in ataxic dysarthria in terms of physiological, perceptual and acoustic characteristics. However, these outcomes have not been fully utilised in the development and sequencing of treatment tasks for ataxic dysarthria. One of the limitations here is the fact that we do not know to what extent ataxic dysarthric speakers should modify their prosodic features (duration, intensity, and fundamental frequency) to improve speech intelligibility when conveying emphasis or stress.

Moreover, research by Patel in 2002 [6] revealed that although dysarthric speakers have a restricted flexibility in prosody control, they are able to communicate the difference between questions and statements using prosodic cues with an accuracy level of up to 98% [6]. However, in a later study in 2014 [7], the ability of ataxic dysarthric speakers to convey contrastive stress was examined. In this study, it was shown that dysarthric speakers mark stress by de-accentuating unstressed words more often than healthy speakers [7]. Nevertheless, some dysarthria speakers were not able to mark contrastive stress based on the underlying perceptual analysis. Therefore, we need to understand the degree of change, in prosodic cues, needed by ataxic dysarthric speakers to correctly mark stress and improve intelligibility.

In this research work, we intend to build an evidence for the treatment of ataxic dysarthria. This is to be achieved by examining the degree of change ataxic dysarthria patients need to make, in terms of intensity, fundamental frequency, and duration, in order to express contrastive stress effectively to listeners. This paper presents the effects of modifying prosodic features; intensity, fundamental frequency and duration, incrementally on the ability of listener's to correctly identify the location of the stressed word within utterances.

The methodology, together with the nature of participants involved, will be described in section 2. The results and outcomes will be presented in section 3. Conclusions and recommendations for possible future research relevant to this study will be discussed in the last section of this paper.

2 Methodology

The methodology used in this study is fully discussed in this chapter. This includes the overview of the participants (both speakers and listeners), initial study, pitch modifications, intensity modifications, duration modifications and the listening experiment.

2.1 Participants

The participants in this investigation include 10 speakers with ataxic dysarthria consisting of 5 males and 5 females as illustrated in Table 1. In addition to the ataxic dysarthric (AT) speakers, 10 healthy control (HC) speakers were also recruited. These healthy control (HC) speakers were aged-matched, as well as gender and dialectal background matched, with the AT speakers. These participants are taken from the dysarthria=c speech data set reported by [8]. The participants have no cognitive deficiency neither do they have any visual and hearing impairment. Their severity varied from mild to severe cases. All of them are monolingual native speakers of English.

Participant	Age	Gender	Etiology	Intelligibility Score (%)
AT_01	46	M	CA	26
AT_02	60	F	CA	33
AT_03	28	M	FA	94
AT_04	52	F	CA	75
AT_05	28	F	FA	91
AT_06	65	F	SCA6	42
AT_07	72	M	CA	81
AT_08	51	M	CA	56
AT_09	56	M	SCA8	18
AT_10	57	F	FA	20

CA: cerebellar ataxia of undefined type, FA: Friedreich's ataxia and SCA: spinocerebellar ataxia

Table 1: Details of participants involved in the study

Their intelligibility scores varied from 18 to 91 as shown in Table 1. These intelligibility scores were estimated from the average scores from five trained listeners during a passage reading task [8]. The etiologies of these participants are either cerebellar ataxia (50%), Friedreich's ataxia (30%) or spinocerebellar ataxia (20%). Each participant produced 30 sentences using the 10 Subject-Verb-Object-Adverbial (SVOA) structured sentences across three stress conditions. These stress conditions are stress on the initial (S), medial (O) and final (A) target words respectively. Audio recordings are saved in the format AT_XX_YY_ZZ where XX is the participant's number (from 01 to 10), YY is the sentence number (from 01 to 10) and ZZ is the stress position (01-initial 02-medial or 03-final position).

In addition, 50 listeners were recruited for the perceptual experiment. These listeners are untrained and do not have any hearing or speech impairment. They are all native speakers of English and mainly university students aged between 18 and 50 years old. Their suitability for the study was tested by engaging them in a practice experiment where each participant is required attain more than 80% accuracy in a stress identification task in normal speech. Listeners with less than 80% stress identification accuracy were excluded from the study. In terms of sample size, the total number of audio samples in the first listening experiment is 212 and that of the second experiment is 259 audio samples.

2.2 Initial Study

A prior study was carried out in [7] on the dataset. This study involved a perception experiment using seven untrained listeners who are native speakers of English and do not have any hearing impairment. During the perception experiment, utterances where more than 60% of the listeners could not locate the target word, were identified. These included utterances where no stress has been placed on any of the words, requiring amplification (AMP) and utterances where the AT speakers produced inappropriate pitch contours (IPC) during the stress marking task. An example of an IPC is when a speaker has stressed every word in the sentence rather than stressing a single target. These identified utterances (both AMP and IPC) formed the baseline (focus sentences) for our study in this paper.

2.3 Pitch Modification

As stated in 2.2, two categories of focus sentences were identified, namely: AMP and IPC. Therefore, two distinct pitch modifications techniques were implemented based on the category of the focus sentence. Pitch incremental modifications were carried out on all the focus sentences (AMP and IPC) while pitch contour modifications were carried out on IPC sentences only.

2.3.1 Pitch Incremental Modification

To establish a reference point for pitch incremental modifications, audio samples from 10 healthy speakers presented in [7] were examined. They (the HC speakers) were given the same SVOA structured sentences and the average increment in the fundamental frequency (F0) of the target word was estimated for the three sentence conditions. HC speakers mark stress by increasing the F0 before a target word and decreasing the F0 after the target word. As illustrated in Figure 1 the average pre-target increments and post-target decrements are 14% and 30% respectively.

Consequently, the pitch of the target words were increased to a maximum of 30% at an incremental rate of 25%, 50%, 75% and 100% (that is, 0.25 of 30% =7.5%, 0.5 of 30% =15%, 0.75 of 30% =22.5% and 1.00 of 30% = 30% respectively). Praat, a speech processing software was used to modify the pitch incrementally. Figure 1 illustrates pitch incremental modifications carried out on AT_08_04_01. The pitch contours are represented by the blue lines in both plots. The F0 of the target word has been increased by 30% while keeping that of other words the same.

2.3.2 Pitch Contour Modification

Pitch contour modifications, involves the modification of the pitch contours of the IPC sentences. The pitch contour modification was implemented using Praat to match the pitch profile of HC. It is important to note that for all pitch contour modifications carried out in this study, the pitch contours of the target words were not modified in any way. Only the pitch contours of other words were modified. The new signal is stored using the synthesis function in Praat.

An example is illustrated in Figure 2 showing the pitch contour of AT04_06_02 before and after pitch contour modifications. Here, the pitch contour of ‘O’ word in the SVOA structured sentence 6 from speaker AT_04 has been preserved. Whereas the pitch contours of the other words in the sentence have been modified to correspond with expected pitch profile for target position 2. Without increasing the F0 of the target word, we can see that the resulting pitch profile shows the location of the target word. It is important to note that sometimes AT speakers can use the wrong pitch contour (for example, falling F0) within the target word. In this case, a more complex pitch contour modification will be required.

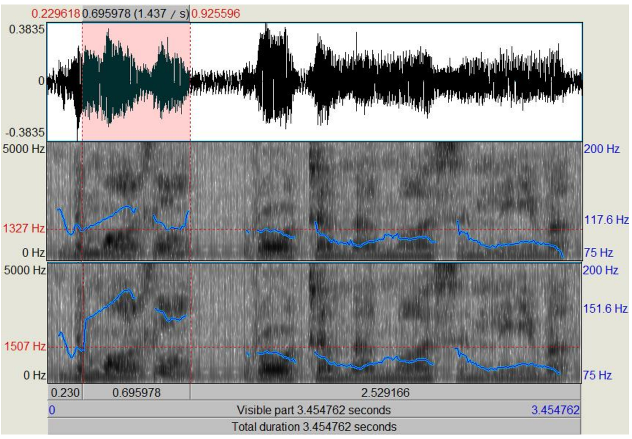

Figure 1: AT_08_04_01 speech before and after 30 % increment in the F0 of the highlighted target word

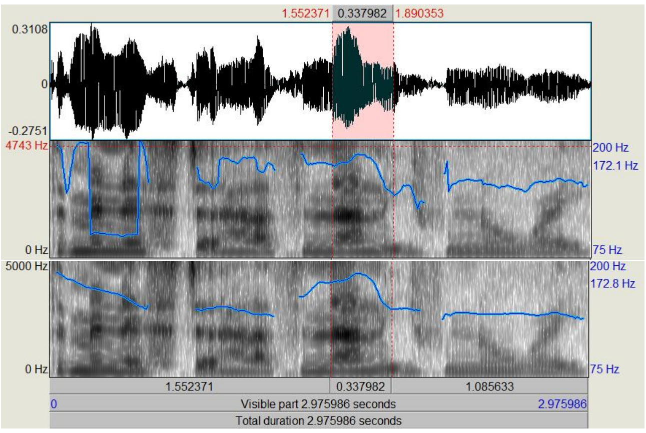

Figure 2: AT04_06_02 before and after pitch contour modification

2.4 Intensity Modification

Apart from increasing F0, research [4, 7] has shown that healthy speakers use increased intensity to mark stress. Healthy speakers increase their intensity just before the target word and decrease the intensity right after the target word. However, for ataxic dysarthric speakers, the variation is intensity is reduced. The relative changes in intensity are dependent on the position of the target word.

In our study, we modified the intensity of the target words in the focus sentences at 4 incremental rates (25%, 50%, 75% and 100%). The software used for these increments was MATLAB R2016b. The intensity increment was achieved by multiplying the amplitude of the target word by the incremental factor as shown in Equation (1).

$$Int_{new} = Int_{old}\,(1 + fac) \qquad (1)$$

where *Int* is the intensity and *fac* is the incremental factor. An example of the intensity modification is illustrated in Figure 3. In this figure, the speaker AT_05 produced sentence 09 with the target word position in the final part of the sentence (03). Looking at both signal, for the original and modified signals, it can be seen that the intensity of the target word has been modified by increasing the amplitude of the highlighted segment of the speech signal. The intensity profiles in dB before intensity modification shows mono loudness. However, after increasing the amplitude of the target word waveform by 100%, the intensity profile shows an emphasis on the target word.

2.5 Duration Modification

Modifying the duration of a target word without altering the intensity or the pitch could be challenging. Over the past few decades, researchers have offered techniques for elongating or shortening of speech signals based on time domain [9] or frequency domain analysis [10]. These techniques included resampling, frequency domain interpolation, and phase vocoder.

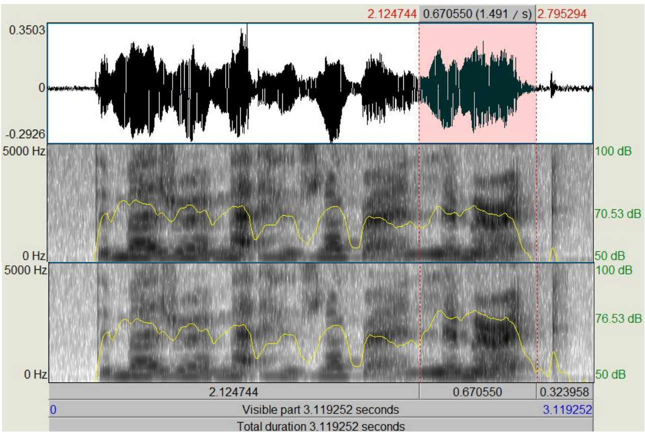

Figure 3: AT_05_09_03 before and after 100% increment in intensity

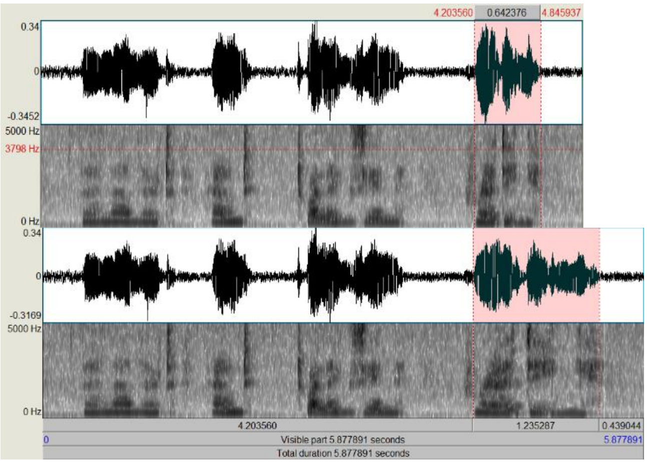

Figure 4: AT_02_03_03 before and after 100% increment in duration

Resampling techniques involve upsampling or downsampling the speech signal in time domain. After which the audio signal is saved at the original sampling frequency. This technique is simple and fast to implement. However, the quality of the signal is compromised leading to modifications in pitch and unnatural sounds. On the other hand, frequency interpolation is a frequency-domain duration modification technique. This technique involves Fourier transform followed by interpolation and inverse Fourier transform. The frequency domain interpolation alters the signal intensity and pitch quality, therefore, resulting in unnatural sounds.

However, for the purpose of this study, we have used the phase vocoder technique. The phase vocoder (PVOC) is a well-known audio synthesis technique used for time dilation and pitch scaling. Time dilation or scaling is achieved by modifying the original short time Fourier transform (STFT) of a signal before performing an inverse short time Fourier transform (ISTFT) on the modified spectrum[11]. Even though the first implementation of PVOC was for a low bit rate speech encoding [11], PVOC has gained high popularity in audio and music processing.

The PVOC is based on the fact that most audio or music signals consist of resonances of sinusoids. The amplitudes and the phases of these sinusoids can be estimated using the STFT function. In the initial application of PVOC, that requires coding and decoding, these amplitudes and phases can be coded (by quantization) and transferred over a channel to a decoder [11]. However, for the purpose of time scaling, the STFT coefficients are modified by keeping the amplitudes the same and modifying the phase so that there are more or fewer oscillation cycles in each frequency band [12]. After which an ISTFT is performed on the modified STFT coefficients.

In our study, a phase interpolation based PVOC was used to modify the duration of the target words. The phase interpolation guarantees horizontal phase coherence; meaning that consecutive frames will overlap coherently [13]. The audio signal is represented as a summation of sinusoids with time-varying amplitudes and instantaneous phases. The PVOC modifies the STFT of the sinusoidal signal by unwrapping the phases of the STFT coefficients. This is achieved by using the increment in phase between two successive frames to estimate the instantaneous frequency of close sinusoid in individual channels [13]. This is a standard time scaling technique and its application in our study is illustrated in Figure 4.

2.6 Listening Experiment

The listening experiment consisted of two stages. The first stage involved individual manipulation of intensity, duration, and the fundamental frequency of target words in both AMP and IPC utterances and pitch contour modification in IPC utterances. However, in the second experiment, individual amplification are combined, that is 2 or 3 amplifications carried out on a single audio sample. In addition, the effects of combining F0, duration and intensity amplification are also examined. The listening experiment set up is illustrated in Table 2. The stress positions (initial, medial and final) are represented by T1, T2 and T3 respectively.

No	**Modifications**			
1	F0	Intensity	Duration	Pitch Contour
2	F0& intensity	F0& duration	Intensity &duration	F0, intensity & duration
	Pause before target		Pause after target	

Table 2: Modifications for perceptual experiments

3 Results and Discussion

For AMP utterances, increasing the duration by 25% in T1 utterances, increases the listener accuracy, though an accuracy drop was experienced in T2 and T3 utterances. However, this drop is not significant as it is within the natural variance of the listeners' perception. Increasing the duration further to 50% improves the listener accuracy significantly ($p<0.05$). A further increase in duration beyond 50% does not give a significant improvement in listener accuracy. This is shown in Figure 5: Acoustic Features Amplification in AMP utterances.

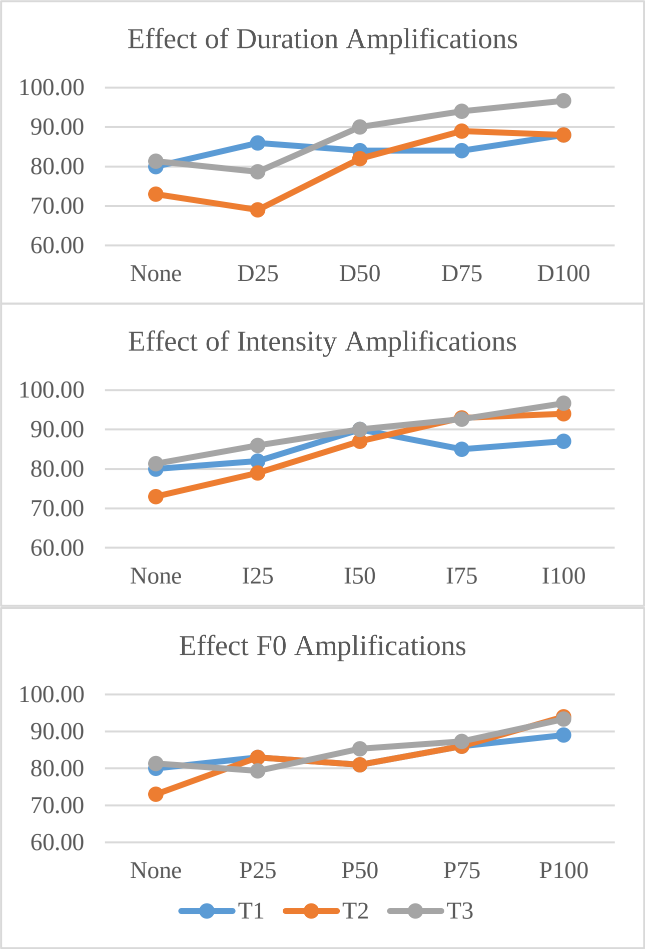

Figure 5: Acoustic Features Amplification in AMP Utterances

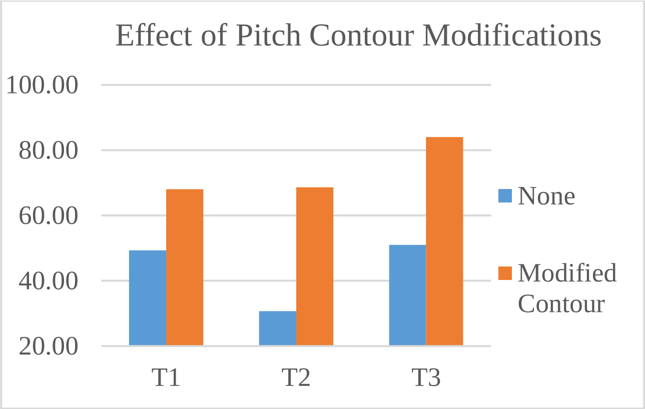

Figure 6: Pitch Modifications in IPC Utterances

Furthermore, increasing the intensity of the target words in AMP utterances improves the listener accuracy as the increment progresses. However, after 50% increase in intensity the improvements in listener accuracy became less significant. In addition, increasing the F0 gave significant improvements from 25% to 50% to75% and to 100%.

Apart from amplification, the effects of pitch contour modifications were also investigated (Figure 6). Improvements in listener accuracy were recorded in all the three sentence conditions. The highest increment was however recorded in T2 utterances and the least increment recorded in T1 utterances.

On the other hand, IPC utterances gave similar significance in listener accuracy improvements. However, the change in listener accuracy in IPC utterances was relatively higher than those experienced in AMP utterances as shown in Figure 7. For example, 50% increment in intensity improved the listener accuracy by 15% in IPC utterances and 10% in AMP utterances.

Moreover, in AMP utterances, the addition of a pause before the target word reduced the listener accuracy by 7% in T2 utterances and 5% in T3 utterances as shown in Figure 8. However, the addition of pauses after the target word increased the listener accuracy by 10% in T1 utterances and 7% in T2 utterances. However, the addition of pauses before the target word in IPC utterance did not have any effect on the listener accuracy while the addition of pauses after the target word improves the listener accuracy by 4% in T1 utterances and 7% in T2 utterances.

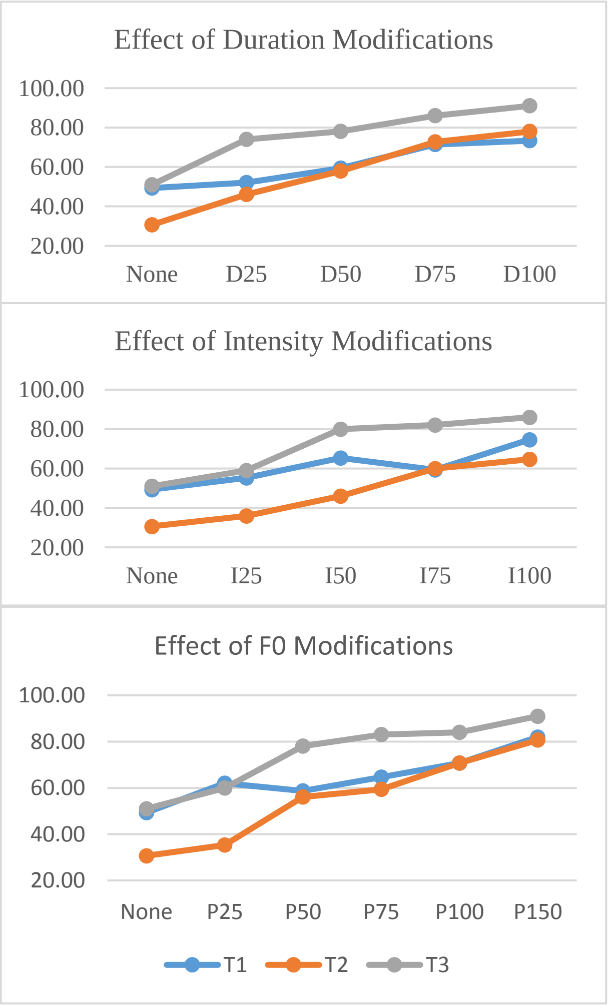

Figure 7: Acoustic Features Amplification in IPC Utterances

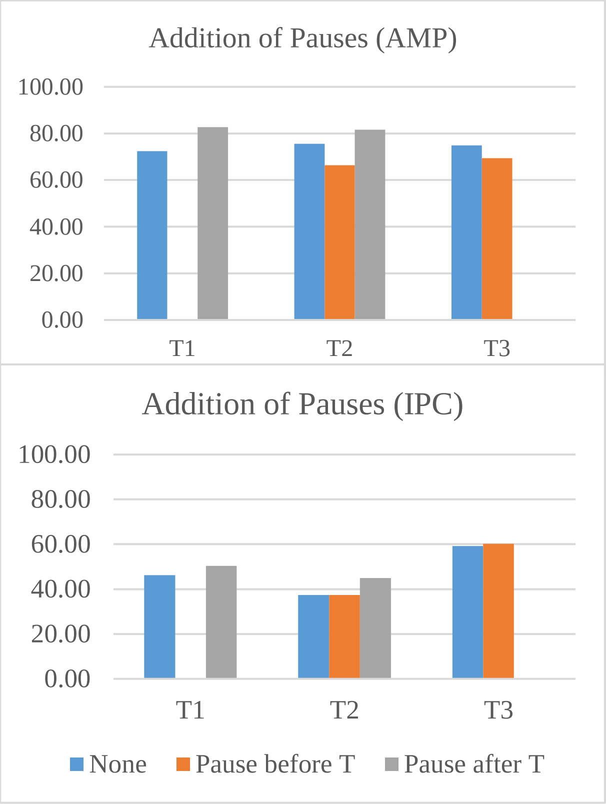

Figure 8: Effects of Addition of Pauses

Consequently, these results imply that 50% increments in duration and intensity are significantly sufficient for improving the listener accuracy. However, a 100% increment in F0 is necessary to significantly improve the listener accuracy. In addition, IPC utterances improved significantly as the acoustic features are increased even though the inappropriate pitch contours were not corrected.

4 Conclusion

The work presented in this paper has shown the effects of increasing intensity, duration or F0 and addition of pauses in dysarthric speech perception. In terms of clinical applications, amplification of any of the 3 stress markers improved the perceptual outcome significantly. Therefore, treatment of dysarthria can be focused on a single area of strength rather than rehabilitating aspects in deficit or combination of parameters. Also, speakers who are unable to amplify their intensity, duration or F0 on target word can add pauses after the target word or change their pitch contour in IPC utterances.

Acknowledgements

This work is funded by The Carnegie Trust for the Universities of Scotland. This is also a collaborative research work between the Electronic and Electrical Engineering and Speech and Language Therapy departments of the University of Strathclyde.

References

[1] J. R. Duffy, *Motor speech disorders: Substrates, differential diagnosis, and management*: Elsevier Health Sciences, 2013.

[2] E. C. Guerra, and D. F. Lovey, "A modern approach to dysarthria classification." pp. 2257-2260 Vol.3.

[3] F. L. Darley, A. E. Aronson, and J. R. Brown, “Differential Diagnostic Patterns of Dysarthria,” *Journal of Speech, Language, and Hearing Research*, vol. 12, no. 2, pp. 246-269, 1969.

[4] R. Patel, and P. Campellone, “Acoustic and perceptual cues to contrastive stress in dysarthria,” *Journal of Speech, Language, and Hearing Research*, vol. 52, no. 1, pp. 206-222, 2009.

[5] K. M. Yorkston, and D. R. Beukelman, “Ataxic DysarthriaTreatment Sequences Based on Intelligibility and Prosodic Considerations,” *Journal of Speech and Hearing Disorders*, vol. 46, no. 4, pp. 398-404, 1981.

[6] R. Patel, “Prosodic control in severe dysarthria: Preserved ability to mark the question-statement contrast,” *Journal of Speech, Language, and Hearing Research*, vol. 45, no. 5, pp. 858, 2002.

[7] A. Lowit, A. Kuschmann, and K. Kavanagh, “Phonological markers of sentence stress in ataxic dysarthria and their relationship to perceptual cues,” *Journal of communication disorders*, vol. 50, pp. 8-18, 2014.

[8] A. Lowit, A. Kuschmann, J. M. MacLeod *et al.*, “Sentence stress in ataxic dysarthria: a perceptual and acoustic study,” *Journal of Medical Speech Language Pathology*, vol. 18, no. 4, pp. 77-82, 2010.

[9] D. Malah, “Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 121-133, 1979.

[10] J. Bonada, "Automatic technique in the frequency domain for near-lossless time-scale modification of audio."

[11] J. L. Flanagan, and R. Golden, “Phase vocoder,” *Bell Labs Technical Journal*, vol. 45, no. 9, pp. 1493-1509, 1966.

[12] A. T. Cemgil, and S. J. Godsill, "Probabilistic phase vocoder and its application to interpolation of missing values in audio signals." pp. 1-4.

[13] J. Laroche, and M. Dolson, “Improved phase vocoder time-scale modification of audio,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323-332, 1999.