Acoustic transformations to improve the intelligibility of dysarthric speech

Frank Rudzicz
University of Toronto, Department of Computer Science
6 King's College Road
Toronto, Ontario, Canada
frank@cs.toronto.edu

Abstract

This paper describes modifications to acoustic speech signals produced by speakers with dysarthria in order to make those utterances more intelligible to typical listeners. These modifications include the correction of tempo, the adjustment of formant frequencies in sonorants, the removal of aberrant voicing, the deletion of phoneme insertion errors, and the replacement of erroneously dropped phonemes. Through simple evaluations of intelligibility with naïve listeners, we show that the correction of phoneme errors results in the greatest increase in intelligibility and is therefore a desirable mechanism for the eventual creation of augmentative application software for individuals with dysarthria.

1 Introduction

Dysarthria is a set of neuromotor disorders that impair the physical production of speech. These impairments reduce the normal control of the primary vocal articulators but do not affect the regular comprehension or production of meaningful, syntactically correct language. For example, damage to the recurrent laryngeal nerve reduces control of vocal fold vibration (i.e., phonation), which can result in aberrant voicing. Inadequate control of soft palate movement caused by disruption of the vagus cranial nerve may lead to a disproportionate amount of air being released through the nose during speech (i.e., hypernasality). The lack of articulatory control also leads to various involuntary non-speech sounds including velopharyngeal or glottal noise (Rosen and Yampolsky, 2000). More commonly, a lack of tongue and lip dexterity often produces heavily slurred speech and a more diffuse and less differentiable vowel target space (Kent and Rosen, 2004).

The neurological damage that causes dysarthria usually affects other physical activity as well which can have a drastically adverse affect on mobility and computer interaction. For instance, severely dysarthric speakers are 150 to 300 times slower than typical users in keyboard interaction (Hosom et al., 2003; Hux et al., 2000). However, since dysarthric speech is often only 10 to 17 times slower than that of typical speakers (Patel, 1998), speech is a viable input modality for computer-assisted interaction.

Consider a dysarthric individual who must travel into a city by public transportation. This might involve purchasing tickets, asking for directions, or indicating intentions to fellow passengers, all within a noisy and crowded environment. A personal portable communication device in this scenario (either hand-held or attached to a wheelchair) would transform relatively unintelligible speech spoken into a microphone to make it more intelligible before being played over a set of speakers. Such a system could facilitate interaction and overcome difficult or failed attempts at communication in daily life.

We propose a system that avoids drawbacks of other voice-output communication aids that output only synthetic speech. Before software for such a device is designed, our goal is to establish and evaluate a set of modifications to dysarthric speech to produce a more intelligible equivalent. Understanding the utility of each of these techniques will be crucial to effectively designing the proposed system.

2 Background and related work

Hawley et al. (2007) described an experiment in which 8 dysarthric individuals (with either cerebral palsy or multiple sclerosis) controlled non-critical devices in their home (e.g., TV) with automatic speech recognition. Command vocabularies consisted of very simple phrases (e.g., "*TV channel up*", "*Radio volume down*") and feedback was provided to the user either by visual displays or by auditory cues. This speech-based environmental control was compared with a 'scanning' interface in which a button is physically pressed to iteratively cycle through a list of alternative commands, words, or phrases. While the speech interface made more errors (between 90.8% and 100% accuracy after training) than the scanning interface (100% accuracy), the former was significantly faster (7.7s vs 16.9s, on average). Participants commented that speech was significantly less tiring than the scanning interface, and just as subjectively appealing (Hawley et al., 2007). Similar results were obtained in other comparisons of speech and scanning interfaces (Havstam, Buchholz, and Hartelius, 2003), and command-and-control systems (Green et al., 2003). Speech is a desirable method of expression for individuals with dysarthria. There are many augmentative communication devices that employ synthetic text-to-speech in which messages can be written on a specialized keyboard or played back from a repository of pre-recorded phrases (Messina and Messina, 2007). This basic system architecture can be modified to allow for the replacement of textual input with spoken input. However, such a scenario would involve some degree of automatic speech recognition, which is still susceptible to fault despite recent advances (Rudzicz, 2011). Moreover, the type of synthetic speech output produced by such systems often lacks a sufficient degree of individual affectation or natural expression that one might expect in typical human speech (Kain et al., 2007). The use of prosody to convey personal information such as one's emotional state is generally not supported by such systems but is nevertheless a key part of a general communicative ability.

Transforming one's speech in a way that preserves the natural prosody will similarly also preserve extra-linguistic information such as emotions, and is therefore a pertinent response to the limitations of current technology. Kain et al. (2007) proposed the voice transformation system shown in figure 1 which produced output speech by concatenating together original unvoiced segments with synthesized voiced segments that consisted of a superposition of the original high-bandwidth signal with synthesized low-bandwidth formants. These synthesized formants were produced by modifications to input energy, pitch generation, and formant modifications. Modifications to energy and formants were performed by Gaussian mixture mapping, as described below, in which learned relationships between dysarthric and target acoustics were used to produce output closer to the target space. This process was intended to be automated, but Kain et al. (2007) performed extensive hand-tuning and manually identified formants in the input. This will obviously be impossible in a real-time system, but these processes can to some extent be automated. For example, voicing boundaries can be identified by the weighted combination of various acoustic features (e.g., energy, zero-crossing rate) (Kida and Kawahara, 2005; Hess, 2008), and formants can be identified by the Burg algorithm (Press et al., 1992) or through simple linear predictive analysis with continuity constraints on the identified resonances between adjacent frames (O'Shaughnessy, 2008).

Spectral modifications traditionally involve filtering or amplification methods such as spectral subtraction or harmonic filtering (O'Shaughnessy, 2000), but these are not useful for dealing with more serious mispronounciations (e.g., $/t/$ for $/n/$). Hosom et al. (2003) showed that Gaussian mixture mapping can be used to transform audio from one set of spectral acoustic features to another. During analysis, context-independent frames of speech are analyzed for bark-scaled energy and their 24^{th} order cepstral coefficients.

For synthesis, a cepstral analysis approximates the original spectrum, and a high-order linear predictive filter is applied to each frame, and excited by impulses or white noise (for voiced and unvoiced segments). Hosom et al. (2003) showed that given 99% human accuracy in recognizing normal speech data, this method of reconstruction gave 93% accuracy on the same data. They then trained a transformative model between dysarthric and regular speech

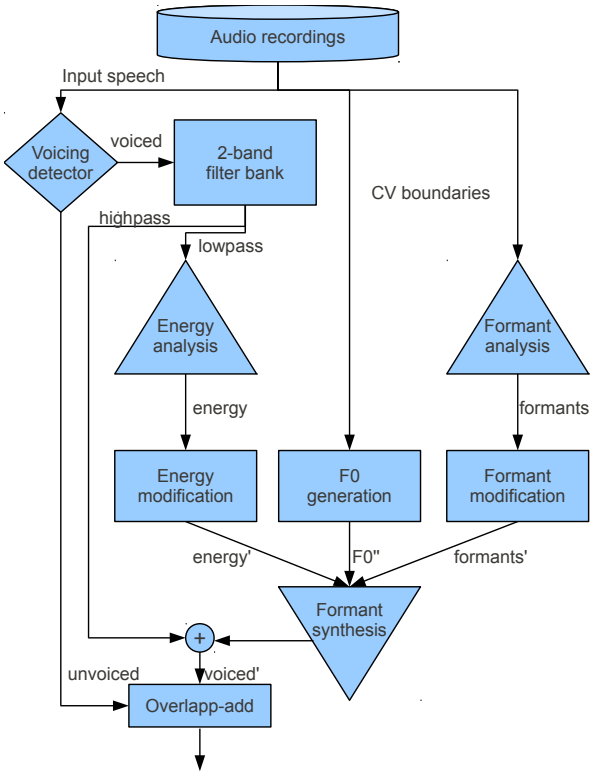


Figure 1: Voice transformation system proposed by Kain et al. (2007).

using aligned, phoneme-annotated, and orthographically identical sentences spoken by dysarthric and regular speakers, and a Gaussian Mixture Model (GMM) to model the probability distribution of the dysarthric source spectral features x as the sum of D normal distributions with mean vector μ, diagonal covariance matrix Σ, and prior probability α:

$$p(x) = \sum_{d=1}^{D} \alpha_d \mathbf{N}\left(x; \mu_d, \Sigma_d\right). \quad (1)$$

The GMM parameters were trained in an unsupervised mode using the expectation-maximization algorithm and 1, 2, 4, 8, and 16 mixture components, with $D = 4$ apparently being optimal. A probabilistic least-squares regression mapped the source features x onto the target (regular speaker) features y, producing the model $W_d(x) + b_d$ for each class, and a simple spectral distortion is performed to produce regularized versions of dysarthic speech $\hat{y}$:

$$\hat{y}(x) = \sum_{d=1}^{D} h_d(x)\left(W_d(x) + b_d\right) \quad (2)$$

for posterior probabilities $h_d(x)$. This model is interesting in that it explicitly maps the acoustic differences for different features between disordered and regular speech[1]. Reconstructing the dysarthric spectrum in this way to sound more 'typical' while leaving pitch (F_0), timing, and energy characteristics intact resulted in a 59.4% relative error rate reduction (68% to 87% accuracy) among a group of 18 naive human listeners each of whom annotated a total of 206 dysarthric test words (Hosom et al., 2003).

3 The TORGOMorph transformations

TORGOMorph encapsulates of a number of transformations of the acoustics uttered by speakers with dysarthria. Each modification is implemented in reaction to a particular effect of dysarthria on intelligibility as determined by observations on the TORGO database of dysarthric speech (Rudzicz, Namasivayam, and Wolff, 2011). Currently, these modifications are uniformly preceded by noise reduction using spectral subtraction and either phonological or phonemic annotations. This latter step is currently necessary, since certain modifications require either knowledge of the manner of articulation or the identities of the vowel segments, as explained below. The purpose of this exercise is to determine which modifications result in the most significant improvements to intelligibility, so the correct annotation sequence is vital to avoid the introduction of an additional dimension of error. Therefore, the annotations used below are extracted directly from the professional markup in the TORGO database. In practice, however, phonemic annotations determined automatically by speech recognition would be imperfect, which is why investigations of this type often forgo that automation altogether (e.g., see Kain et al. (2007)). Possible alternatives to full ASR are discussed in section 5.

In some cases, the dysarthric speech must be compared or supplemented with another vocal source. Here, we synthesize segments of speech using a text-to-speech application developed by Black and Lenzo (2004). This system is based on the University of Edinburgh's Festival tool and synthesizes phonemes using a standard method based on lin-

[1]This model can also be used to measure the difference between any two types of speech.

ear predictive coding with a pronunciation lexicon and part-of-speech tagger that assists in the selection of intonation parameters (Taylor, Black, and Caley, 1998). This system is invoked by providing the expected text uttered by the dysarthic speaker. In order to properly combine this purely synthetic signal and the original waveforms we require identical sampling rates, so we resample the former by a rational factor using a polyphase filter with low-pass filtering to avoid aliasing (Hayes, 1999). Since the discrete phoneme sequences themselves can differ, we find an ideal alignment between the two by the Levenshtein algorithm (Levenshtein, 1966), which provides the total number of insertion, deletion, and substitution errors.

The following sections detail the components of TORGOMorph, which is outlined in figure 2. These components allow for a cascade of one transformation followed by another, although we can also perform these steps independently to isolate their effects. In all cases, the spectrogram is derived with the fast Fourier transform given 2048 bins on the range of 0–5 kHz. Voicing boundaries are extracted in a unidimensional vector aligned with the spectrogram using the method of Kida and Kawahara (2005) which uses GMMs trained with zero-crossing rate, amplitude, and the spectrum as input parameters. A pitch (F_0) contour is also extracted from the source by the method proposed by Kawahara et al. (2005), which uses a Viterbi-like potential decoding of F_0 traces described by cepstral and temporal features. That work showed an error rate of less than 0.14% in estimating F_0 contours as compared with simultaneously-recorded electroglottograph data. These contours are not in general modified by the methods proposed below, since Kain et al. (2007) showed that using original F_0 results in the highest intelligibility among alternative systems. Over a few segments, however, these contours can sometimes be decimated in time during the modification proposed in section 3.3 and in some cases removed entirely (along with all other acoustics) in the modification proposed in section 3.2.

3.1 High-pass filter on unvoiced consonants

The first acoustic modification is based on the observation that unvoiced consonants are improperly voiced in up to 18.7% of plosives (e.g. */d/* for */t/*) and up to 8.5% of fricatives (e.g. */v/* for */f/*) in dysarthric speech in the TORGO database. Voiced consonants are typically differentiated from their unvoiced counterparts by the presence of the *voice bar*, which is a concentration of energy below 150 Hz indicative of vocal fold vibration that often persists throughout the consonant or during the closure before a plosive (Stevens, 1998). Empirical analysis of TORGO data suggests that for at least two male dysarthric speakers this voice bar extends considerably higher, up to 250 Hz.

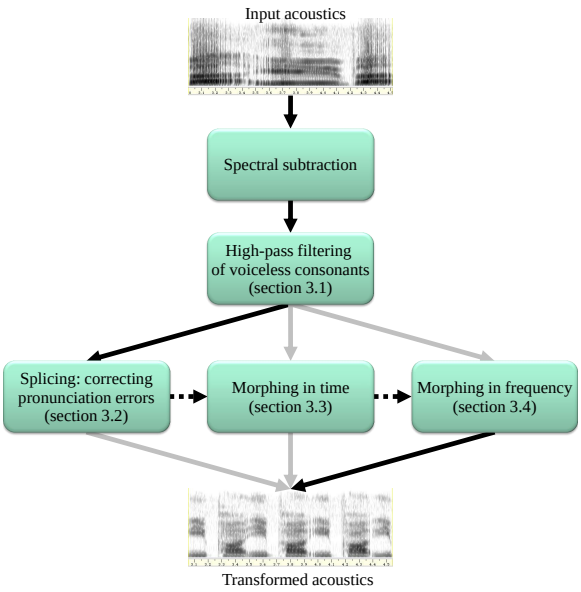


Figure 2: Outline of the TORGOMorph system. The black path indicates the cascade to be used in practice. Solid arrows indicate paths taken during evaluation.

In order to correct these mispronunciations, the voice bar is filtered out of all acoustic sub-sequences annotated as unvoiced consonants. For this task we use a high-pass Butterworth filter, which is "maximally flat" in the passband[2] and monotonic in magnitude in the frequency domain (Butterworth, 1930). Here, this filter is computed on a normalized frequency range respecting the Nyquist frequency, so that if a waveform's sampling rate is 16 kHz, the normalized cutoff frequency for this component is $f^*_{Norm} = 250/(1.6 \times 10^4/2) = 3.125 \times 10^{-2}$. The Butterworth filter is an all-pole transfer function between signals, and we use the 10^{th}-order low-pass

[2]The passband is the frequency range in which the component magnitudes in the original signal should not be changed.

Butterworth filter whose magnitude response is

$$|\mathcal{B}(z;10)|^2 = |H(z;10)|^2 = \frac{1}{1+\left(jz/jz^*_{Norm}\right)^{2\times 10}} \tag{3}$$

where z is the complex frequency in polar coordinates and z^*_{Norm} is the cutoff frequency in that domain (Hayes, 1999). This allows the transfer function

$$\mathcal{B}(z;10) = H(z;10) = \frac{1}{1+z^{10}+\sum_{i=1}^{10} c_i z^{10-i}} \tag{4}$$

whose poles occur at known symmetric intervals around the unit complex-domain circle (Butterworth, 1930). These poles are then transformed by the Matlab function `zp2ss`, which produces the state-space coefficients α_i and β_i that describe the output signal resulting from applying the low-pass Butterworth filter to the discrete signal $x[n]$. These coefficients are further converted by

$$\begin{aligned} \vec{a} &= z^*_{Norm}\vec{\alpha}^{-1} \\ \vec{b} &= -z^*_{Norm}\left(\vec{\alpha}^{-1}\vec{\beta}\right) \end{aligned} \tag{5}$$

giving the high-pass Butterworth filter with the same cutoff frequency of z^*_{Norm}. This continuous system is converted to the discrete equivalent through the impulse-invariant discretization method and is implemented by the difference equation

$$y[n] = \sum_{k=1}^{10} a_k y[n-k] + \sum_{k=0}^{10} b_k x[n-k]. \tag{6}$$

As previously mentioned, this equation is applied to each acoustic sub-sequence annotated as unvoiced consonants, thereby smoothly removing the energy below 250 Hz.

3.2 Splicing: correcting dropped and inserted phoneme errors

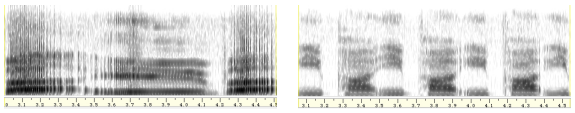
The Levenshtein algorithm finds a best possible alignment of the phoneme sequence in actually uttered speech and the expected phoneme sequence, given the known word sequence. Isolating phoneme insertions and deletions are therefore a simple matter of iteratively adjusting the source speech according to that alignment. There are two cases where action is required:

insertion error In this case a phoneme is present where it ought not be. In the TORGO database, these insertion errors tend to be repetitions of phonemes occurring in the first syllable of a word, according to the International Speech Lexicon Dictionary (Hasegawa-Johnson and Fleck, 2007). When an insertion error is identified the entire associated segment of the signal is simply removed. In the case that the associated segment is not surrounded by silence, adjacent phonemes can be merged together with time-domain pitch-synchronous overlap-add (Moulines and Charpentier, 1990).

deletion error The vast majority of accidentally deleted phonemes in the TORGO database are fricatives, affricates, and plosives. Often, these involve not properly pluralizing nouns (e.g., *book* instead of *books*). Given their high preponderance of error, these phonemes are the only ones we insert into the dysarthric source speech. Specifically, when the deletion of a phoneme is recognized with the Levenshtein algorithm, we simply extract the associated segment from the aligned synthesized speech and insert it into the appropriate spot in the dysarthric speech. For all unvoiced fricatives, affricates, and plosives no further action is required. When these phonemes are voiced, however, we first extract and remove the F_0 curve from the synthetic speech, linearly interpolate the F_0 curve from adjacent phonemes in the source dysarthric speech, and resynthesize with the synthetic spectrum and interpolated F_0. If interpolation is not possible (e.g., the synthetic voiced phoneme is to be inserted beside an unvoiced phoneme), we simply generate a flat F_0 equal to the nearest natural F_0 curve.

3.3 Morphing in time

Figure 3 exemplifies that vowels uttered by dysarthric speakers are significantly slower than those uttered by typical speakers. In fact, sonorants can be twice as long in dysarthric speech, on average (Rudzicz, Namasivayam, and Wolff, 2011). In this modification, phoneme sequences identified as sonorant are simply contracted in time in order to be equal in extent to the greater of half their original



(a) (b)

Figure 3: Repetitions of */iy p ah/* over 1.5s by (a) a male speaker with athetoid CP, and (b) a female control in the TORGO database. Dysarthric speech is notably slower and more strained than regular speech.

length or the equivalent synthetic phoneme's length. In all cases this involved shortening the dysarthric source sonorant.

Since we wish to contract the length of a signal segment here without affecting its pitch or frequency characteristics, we use a phase vocoder based on digital short-time Fourier analysis (Portnoff, 1976). Here, Hamming-windowed segments of the source phoneme are analyzed with a z-transform giving both frequency and phase estimates for up to 2048 frequency bands. During pitch-preserving time-scaled warping, we specify the magnitude spectrum directly from the input magnitude spectrum with phase values chosen to ensure continuity (Sethares, 2007). Specifically, for the frequency band at frequency F and frames j and $k > j$ in the modified spectrogram, the phase θ is predicted by

$$\theta_k^{(F)} = \theta_j^{(F)} + 2\pi F(j-k). \tag{7}$$

In our case the discrete warping of the spectrogram involves simple decimation by a constant factor. The spectrogram is then converted into a time-domain signal modified in tempo but not in pitch relative to the original phoneme segment. This conversion is accomplished simply through the inverse Fourier transform.

3.4 Morphing in frequency

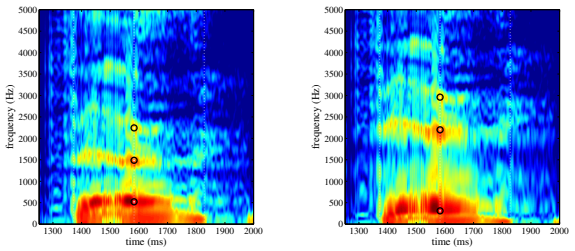
Formant trajectories inform the listener as to the identities of vowels, but the vowel space of dysarthric speakers tends to be constrained (Kain et al., 2007). In order to improve a listener's ability to differentiate between the vowels, this modification component identifies formant trajectories in the acoustics and modifies these according to the known vowel identity of a segment. Here, formants are identified with a 14^{th}-order linear-predictive coder with continuity constraints on the identified resonances between adjacent frames (Snell and Milinazzo, 1993; O'Shaughnessy, 2008). Bandwidths are determined by the negative natural logarithm of the pole magnitude, as implemented in the STRAIGHT analysis system (Banno et al., 2007; Kawahara, 2006).

For each identified vowel in the dysarthric speech[3], formant candidates are identified at each frame in time up to 5 kHz. Only those time frames having at least 3 such candidates within 250 Hz of expected values are considered. The expected values of formants are derived from analyses performed by Allen et al. (1987). Given these subsets of candidate time frames in the vowel, the one having the highest spectral energy within the middle 50% of the length of the vowel is established as the *anchor position*, and the three formant candidates within the expected ranges are established as the *anchor frequencies* for formants F_1 to F_3. If more than one formant candidate falls within expected ranges, the one with the lowest bandwidth becomes the anchor frequency.

Given identified anchor points and target sonorant-specific frequencies and bandwidths, there are several methods to modify the spectrum. The most common may be to learn a statistical conversion function based on Gaussian mixture mapping, as described earlier, typically preceded by alignment of sequences using dynamic time warping (Stylianou, 2008). Here, we use the STRAIGHT morphing implemented by Kawahara and Matsui (2003), among others. The transformation of a frame of speech x_A for speaker A is performed with a multivariate frequency-transformation function $T_{A\beta}$ given known targets β using

$$\begin{aligned} T_{A\beta}(x_A) &= \int_0^{x_A} \exp\left(\log\left(\frac{\delta T_{A\beta}(\lambda)}{\delta\lambda}\right)\right)\delta\lambda \\ &= \int_0^{x_A} \exp\left((1-r)\log\left(\frac{\delta T_{AA}(\lambda)}{\delta\lambda}\right)\right. \\ &\quad \left. + r\log\left(\frac{\delta T_{A\beta}(\lambda)}{\delta\lambda}\right)\right)\delta\lambda \\ &= \int_0^{x_A} \left(\frac{\delta T_{A\beta}(\lambda)}{\delta\lambda}\right)^r \delta\lambda, \end{aligned} \tag{8}$$

[3] Accidentally inserted vowels are also included here, unless previously removed by the splicing technique in section 3.2.



(a) (b)

Figure 4: Spectrograms for (a) the dysarthric original and (b) the frequency-modified renditions of the word *fear*. Circles represent indicative formant locations.

where λ is the frame-based time dimension and where $0 \leq r \leq 1$ is an interpolative rate at which to perform morphing (i.e., $r = 1$ implies complete conversion of the parameters of speaker A to parameter set β and $r = 0$ implies no conversion.) (Kawahara et al., 2009). An example of the results of this morphing technique is shown in figure 4 in which the three identified formants are shifted to their expected frequencies.

This method tracks formants and warps the frequency space automatically, whereas Kain et al. (2007) perform these functions manually. A future implementation may use Kalman filters to reduce the noise inherent in trajectory tracking. Such an approach has shown significant improvements in formant tracking, especially for F_1 (Yan et al., 2007).

4 Intelligibility experiments with TORGOMorph

The intelligibility of both purely synthetic and modified speech signals can be measured objectively by simply having a set of participants transcribe what they hear from a selection of word, phrase, or sentence prompts (Spiegel et al., 1990), although no single standard has emerged as pre-eminent (Schroeter, 2008). Hustad (2006) suggested that orthographic transcriptions provide a more accurate predictor of intelligibility among dysarthric speakers than the more subjective estimates used in clinical settings, e.g., Enderby (1983). That study had 80 listeners who transcribed audio (which is an atypically large group for this task) and showed that intelligibility increased from 61.9% given only acoustic stimuli to 66.75% given audiovisual stimuli on the transcription task in normal speech. In the current work, we modify only the acoustics of dysarthric speech; however future work might consider how to prompt listeners in a more multimodal context.

In order to gauge the intelligibility of our modifications, we designed a simple experiment in which human listeners attempt to identify words in sentence-level utterances under a number of acoustic scenarios. Sentences are either uttered by a speaker with dysarthria, modified from their original source acoustics, or manufactured by a text-to-speech synthesizer. Each participant is seated at a personal computer with a simple graphical user interface with a button which plays or replays the audio (up to 5 times), a text box in which to write responses, and a second button to submit those responses. Audio is played over a pair of headphones. The participants are told to only transcribe the words with which they are reasonably confident and to ignore those that they cannot discern. They are also informed that the sentences are grammatically correct but not necessarily semantically coherent, and that there is no profanity. Each participant listens to 20 sentences selected at random with the constraints that at least two utterances are taken from each category of audio, described below, and that at least five utterances are also provided to another listener, in order to evaluate inter-annotator agreement. Participants are self-selected to have no extensive prior experience in speaking with individuals with dysarthria, in order to reflect the general population. Although dysarthric utterances are likely to be contextualized within meaningful conversations in real-world situations, such pragmatic aspects of discourse are not considered here in order to concentrate on acoustic effects alone. No cues as to the topic or semantic context of the sentences are given, as there is no evidence that such aids to comprehension affect intelligibility (Hustad and Beukelman, 2002). In this study we use sentence-level utterances uttered by male speakers from the TORGO database.

Baseline performance is measured on the original dysarthric speech. Two other systems are used for reference:

Synthetic Word sequences are produced by the Cepstral commercial text-to-speech system using the U.S. English voice 'David'. This sys-

tem is based on Festival in almost every respect, including its use of linguistic pre-processing (e.g., part-of-speech tagging) and rule-based generation (Taylor, Black, and Caley, 1998). This approach has the advantage that every aspect of the synthesized speech (e.g., the word sequence) can be controlled although here, as in practice, synthesized speech will not mimic the user's own acoustic patterns, and will often sound more 'mechanical' due to artificial prosody (Black and Lenzo, 2007).

GMM This system uses the Gaussian mixture mapping type of modification suggested by Toda, Black, and Tokuda (2005) and Kain et al. (2007). Here, we use the FestVox implementation of this algorithm, which includes pitch extraction, some phonological knowledge (Toth and Black, 2005), and a method for resynthesis. Parameters for this model are trained by the FestVox system using a standard expectation-maximization approach with 24^{th}-order cepstral coefficients and 4 Gaussian components. The training set consists of all vowels uttered by a male speaker in the TORGO database and their synthetic realizations produced by the method above.

Performance is evaluated on the three other acoustic transformations, namely those described in sections 3.2, 3.3, and 3.4 above. Tables 1 and 2 respectively show the percentage of words and phonemes correctly identified by each listener relative to the expected word sequence under each acoustic condition. In each case, annotator transcriptions were aligned with the 'true' or expected sequences using the Levenshtein algorithm described in section 3. Plural forms of singular words, for example, are considered incorrect in word alignment although one obvious spelling mistake (i.e., 'skilfully') is corrected before evaluation. Words are split into component phonemes according to the CMU dictionary, with words having multiple pronunciations given the first decomposition therein.

In these experiments there is not enough data from which to make definitive claims of statistical significance, but it is clear that the purely synthetic speech has a far greater intelligibility than other approaches, more than doubling the average accuracy of the TORGOMorph modifications. The GMM transformation method proposed by Kain et al. (2007) gave poor performance, although our experiments are distinguished from theirs in that our formant traces are detected automatically, rather than by hand. The relative success of the synthetic approach is not an argument against the type of modifications proposed here and by Kain et al. (2007), since our aim is to avoid the use of impersonal and invariant utterances. Indeed, future study in this area should incorporate subjective measures of 'naturalness'. Further uses of acoustic modifications not attainable by text-to-speech synthesis are discussed in section 5.

	Orig.	GMM	Synth.	Splice	Time	Freq.
L01	22.1	15.6	82.0	40.2	34.7	35.2
L02	27.8	12.2	75.5	44.9	39.4	33.8
L03	38.3	14.8	76.3	37.5	12.9	21.4
L04	24.7	10.8	72.1	32.6	22.2	18.4
Avg.	28.2	13.6	76.5	38.8	27.3	27.2

Table 1: Percentage of *words* correctly identified by each listener (L0*) relative to the expected sequence. Sections 3.2, 3.3, and 3.4 discuss the 'Splice', 'Time', and 'Freq.' techniques, respectively.

	Orig.	GMM	Synth.	Splice	Time	Freq.
L01	52.0	43.1	98.2	64.7	47.8	55.1
L02	57.8	38.2	92.9	68.9	50.6	53.3
L03	50.1	41.4	96.8	57.1	30.7	46.7
L04	51.6	33.8	88.7	51.9	43.2	45.0
Avg.	52.9	39.1	94.2	60.7	43.1	50.0

Table 2: Percentage of *phonemes* correctly identified by each listener relative to the expected sequence. Sections 3.2, 3.3, and 3.4 discuss the 'Splice', 'Time', and 'Freq.' techniques, respectively.

In all cases, the splicing technique of removing accidentally inserted phonemes and inserting missing ones gives the highest intelligibility relative to all acoustic transformation methods. Although more study is required, this result emphasizes the importance of lexically correct phoneme sequences. In the word-recognition experiment, there are an average of 5.2 substitution errors per sentence in the unmodified dysarthric speech against 2.75 in the synthetic speech. There are also 2.6 substitution errors on average per sentence for the speech modified in frequency, but 3.1 deletion errors, on average, against 0.24 in synthetic speech. No correlation is found be-

tween the 'loudness' of the speech (determined by the overall energy in the sonorants) and intelligibility results, although this might change with the acquisition of more data. Neel (2009), for instance, found that loud or amplified speech from individuals with Parkinson's disease was more intelligible to human listeners than quieter speech.

Our results are comparable in many respects to the experiments of Kain et al. (2007), although they only looked at simple consonant-vowel-consonant stimuli. Their results showed an average of 92% correct synthetic vowel recognition (compared with 94.2% phoneme recognition in table 2) and 48% correct dysarthric vowel recognition (compared with 52.9% in table 2). Our results, however, show that modified timing and modified frequencies do not actually benefit intelligibility in either the word or phoneme cases. This disparity may in part be due to the fact that our stimuli are much more complex (quicker sentences do not necessarily improve intelligibility).

5 Discussion

This work represents an inaugural step towards speech modification systems for human-human and human-computer interaction. Tolba and Torgoman (2009) claimed that significant improvements in automatic recognition of dysarthric speech are attainable by modifying formants F_1 and F_2 to be more similar to expected values. In that study, formants were identified using standard linear predictive coding techniques, although no information was provided as to how these formants were modified nor how their targets were determined. However, they claimed that modified dysarthric speech resulted in 'recognition rates' (by which they presumably meant word-accuracy) of 71.4% in the HTK speech recognition system, as compared with 28% on the unmodified dysarthric speech from 7 individuals. The results in section 4 show that human listeners are more likely to correctly identify utterances in which phoneme insertion and deletion errors are corrected than those in which formant frequencies are adjusted. Therefore, one might hypothesize that such pre-processing might provide even greater gains than those reported by Tolba and Torgoman (2009). Ongoing work ought to confirm or deny this hypothesis.

A prototypical client-based application based on our research for unrestricted speech transformation of novel sentences is currently in development. Such work will involve improving factors such as accuracy and accessibility for individuals whose neuromotor disabilities limit the use of modern speech recognition, and for whom alternative interaction modalities are insufficient. This application is being developed under the assumption that it will be used in a mobile device embeddable within a wheelchair. If word-prediction is to be incorporated, the predicted continuations of uttered sentence fragments can be synthesized without requiring acoustic input.

In practice, the modifications presented here will have to be based on automatically-generated annotations of the source audio. This is especially important to the 'splicing' module in which word-identification is crucial. There are a number of techniques that can be exercised in this area. Czyzewski, Kaczmarek, and Kostek (2003) apply both a variety of neural networks and rough sets to the task of classifying segments of speech according to the presence of stop-gaps, vowel prolongations, and incorrect syllable repetitions. In each case, input includes source waveforms and detected formant frequencies. They found that stop-gaps and vowel prolongations could be detected with up to 97.2% accuracy and that vowel repetitions could be detected with up to 90% accuracy using the rough set method. Accuracy was similar although slightly lower using traditional neural networks (Czyzewski, Kaczmarek, and Kostek, 2003). These results appear generally invariant even under frequency modifications to the source speech. Arbisi-Kelm (2010), for example, suggest that disfluent repetitions can be identified reliably through the use of pitch, duration, and pause detection (with precision up to 93% (Nakatani, 1993)). If more traditional models of speech recognition are to be deployed to identify vowels, the probabilities that they generate across hypothesized words might be used to weight the manner in which acoustic transformations are made.

The use of one's own voice to communicate is a desirable goal, and continuations of this research are therefore focused on the practical aspects of this research towards usable and portable systems.

References

Allen, Jonathan, M. Sharon Hunnicutt, Dennis H. Klatt, Robert C. Armstrong, and David B. Pisoni. 1987. *From text to speech: the MITalk system*. Cambridge University Press, New York, NY, USA.

Arbisi-Kelm, Timothy. 2010. Intonation structure and disfluency detection in stuttering. *Laboratory Phonology 10*, 4:405–432.

Banno, Hideki, Hiroaki Hata, Masanori Morise, Toru Takahashi, Toshio Irino, and Hideki Kawahara. 2007. Implementation of realtime STRAIGHT speech manipulation system: Report on its first implementation. *Acoustical Science and Technology*, 28(3):140–146.

Black, Alan W. and Kevin A. Lenzo. 2004. Multilingual text-to-speech synthesis. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*.

Black, Alan W. and Kevin A. Lenzo. 2007. Building synthetic voices. `http://www.festvox.org/festvox/bsv.ps.gz`.

Butterworth, Stephen. 1930. On the theory of filter amplifiers. *Experimental Wireless and the Wireless Engineer*, 7:536–541.

Czyzewski, Andrzej, Andrzej Kaczmarek, and Bozena Kostek. 2003. Intelligent processing of stuttered speech. *Journal of Intelligent Information Systems*, 21(2):143–171.

Enderby, Pamela M. 1983. *Frenchay Dysarthria Assessment*. College Hill Press.

Green, Phil, James Carmichael, Athanassios Hatzis, Pam Enderby, Mark Hawley, and Mark Parker. 2003. Automatic speech recognition with sparse training data for dysarthric speakers. In *Proceedings of Eurospeech 2003*, pages 1189–1192, Geneva.

Hasegawa-Johnson, Mark and Margaret Fleck. 2007. International Speech Lexicon Project. `http://www.isle.illinois.edu/dict/`.

Havstam, Christina, Margret Buchholz, and Lena Hartelius. 2003. Speech recognition and dysarthria: a single subject study of two individuals with profound impairment of speech and motor control. *Logopedics Phoniatrics Vocology*, 28:81–90(10), August.

Hawley, Mark S., Pam Enderby, Phil Green, Stuart Cunningham, Simon Brownsell, James Carmichael, Mark Parker, Athanassios Hatzis, Peter O'Neill, and Rebecca Palmer. 2007. A speech-controlled environmental control system for people with severe dysarthria. *Medical Engineering & Physics*, 29(5):586–593, June.

Hayes, Monson H. 1999. *Digital Signal Processing*. Schaum's Outlines. McGraw Hill.

Hess, Wolfgang J. 2008. Pitch and voicing determination of speech with an extension toward music signal. In Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, editors, *Speech Processing*. Springer.

Hosom, John-Paul, Alexander B. Kain, Taniya Mishra, Jan P. H. van Santen, Melanie Fried-Oken, and Janice Staehely. 2003. Intelligibility of modifications to dysarthric speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, volume 1, pages 924–927, April.

Hustad, Katherine C. 2006. Estimating the intelligibility of speakers with dysarthria. *Folia Phoniatrica et Logopaedica*, 58(3):217–228.

Hustad, Katherine C. and David R. Beukelman. 2002. Listener comprehension of severely dysarthric speech: Effects of linguistic cues and stimulus cohesion. *Journal of Speech, Language, and Hearing Research*, 45:545–558, June.

Hux, Karen, Joan Rankin-Erickson, Nancy Manasse, and Elizabeth Lauritzen. 2000. Accuracy of three speech recognition systems: Case study of dysarthric speech. *Augmentative and Alternative Communication (AAC)*, 16(3):186 –196, January.

Kain, Alexander B., John-Paul Hosom, Xiaochuan Niu, Jan P.H. van Santen, Melanie Fried-Oken, and Janice Staehely. 2007. Improving the intelligibility of dysarthric speech. *Speech Communication*, 49(9):743–759, September.

Kawahara, H. and H. Matsui. 2003. Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, volume 1, pages I–256 – I–259 vol.1, April.

Kawahara, H., R. Nisimura, T. Irino, M. Morise, T. Takahashi, and H. Banno. 2009. Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, pages 3905–3908, April.

Kawahara, Hideki. 2006. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology*, 27(6):349–353.

Kawahara, Hideki, Alain de Cheveigné, Hideki Banno, Toru Takahashi, and Toshio Irino. 2005. Nearly Defect-Free F0 Trajectory Extraction for Expressive Speech Modifications Based on STRAIGHT. In *Proceedings of INTERSPEECH 2005*, pages 537–540, September.

Kent, Ray D. and Kristin Rosen. 2004. Motor control perspectives on motor speech disorders. In Ben

Maassen, Raymond Kent, Herman Peters, Pascal Van Lieshout, and Wouter Hulstijn, editors, *Speech Motor Control in Normal and Disordered Speech*. Oxford University Press, Oxford, chapter 12, pages 285–311.

Kida, Yusuke and Tatsuya Kawahara. 2005. Voice activity detection based on optimally weighted combination of multiple features. In *Proceedings of INTERSPEECH-2005*, pages 2621–2624.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8):707–710.

Messina, James J. and Constance M. Messina. 2007. Description of AAC devices. `http://www.coping.org/specialneeds/assistech/aacdev.htm`, April.

Moulines, Eric and Francis Charpentier. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467, December.

Nakatani, Christine. 1993. A speech-first model for repair detection and correction. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 46–53.

Neel, Amy T. 2009. Effects of loud and amplified speech on sentence and word intelligibility in parkinson disease. *Journal of Speech, Language, and Hearing Research*, 52:1021–1033, August.

O'Shaughnessy, Douglas. 2000. *Speech Communications – Human and Machine*. IEEE Press, New York, NY, USA.

O'Shaughnessy, Douglas. 2008. Formant estimation and tracking. In Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, editors, *Speech Processing*. Springer.

Patel, Rupal. 1998. Control of prosodic parameters by an individual with severe dysarthria. Technical report, University of Toronto, December.

Portnoff, Michael R. 1976. Implementation of the digital phase vocoder using the fast Fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(3):243–248.

Press, William H., Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 1992. *Numerical Recipes in C: the art of scientific computing*. Cambridge University Press, second edition.

Rosen, Kristin and Sasha Yampolsky. 2000. Automatic speech recognition and a review of its functioning with dysarthric speech. *Augmentative & Alternative Communication*, 16(1):48–60, Jan.

Rudzicz, Frank. 2011. *Production knowledge in the recognition of dysarthric speech*. Ph.D. thesis, University of Toronto, Department of Computer Science.

Rudzicz, Frank, Aravind Kumar Namasivayam, and Talya Wolff. 2011. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation,* (in press).

Schroeter, Juergen. 2008. Basic principles of speech synthesis. In Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, editors, *Speech Processing*. Springer.

Sethares, William Arthur. 2007. *Rhythm and Transforms*. Springer.

Snell, Roy C. and Fausto Milinazzo. 1993. Formant Location from LPC Analysis Data. *IEEE Transactions on Speech and Audio Processing*, 1(2), April.

Spiegel, Murray F., Mary Jo Altom, Marian J. Macchi, and Karen L. Wallace. 1990. Comprehensive assessment of the telephone intelligibility of synthesized and natural speech. *Speech Communication*, 9(4):279 – 291.

Stevens, Kenneth N. 1998. *Acoustic Phonetics*. MIT Press, Cambridge, Massachussetts.

Stylianou, Yannis. 2008. Voice transformation. In Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang, editors, *Speech Processing*. Springer.

Taylor, Paul, Alan W. Black, and Richard Caley. 1998. The architecture of the Festival speech synthesis system. In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, pages 147–151, Jenolan Caves, Australia.

Toda, Tomoki, Alan W. Black, and Keiichi Tokuda. 2005. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. In *Proceedings of the 2005 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, Philadelphia, Pennsylvania.

Tolba, Hesham and Ahmed S. El Torgoman. 2009. Towards the improvement of automatic recognition of dysarthric speech. In *International Conference on Computer Science and Information Technology*, pages 277–281, Los Alamitos, CA, USA. IEEE Computer Society.

Toth, Arthur R. and Alan W. Black. 2005. Cross-speaker articulatory position data for phonetic feature prediction. In *Proceedings of Interspeech 2005*, Lisbon, Portugal.

Yan, Qin, Saeed Vaseghi, Esfandiar Zavarehei, Ben Milner, Jonathan Darch, Paul White, and Ioannis Andrianakis. 2007. Formant tracking linear prediction model usng HMMs and Kalman filters for noisy speech processing. *Computer Speech and Language*, 21:543–561.