



# Adjusting dysarthric speech signals to be more intelligible<sup>☆</sup>

Frank Rudzicz<sup>\*</sup>

*Department of Computer Science, University of Toronto, 10 King's College Road, Room 3302 Toronto, Ontario M5S 3G4, Canada*

Received 29 December 2011; received in revised form 29 October 2012; accepted 23 November 2012

Available online 6 December 2012

## Abstract

This paper presents a system that transforms the speech signals of speakers with physical speech disabilities into a more intelligible form that can be more easily understood by listeners. These transformations are based on the correction of pronunciation errors by the removal of repeated sounds, the insertion of deleted sounds, the devoicing of unvoiced phonemes, the adjustment of the tempo of speech by phase vocoding, and the adjustment of the frequency characteristics of speech by anchor-based morphing of the spectrum. These transformations are based on observations of disabled articulation including improper glottal voicing, lessened tongue movement, and lessened energy produced by the lungs. This system is a substantial step towards full automation in speech transformation without the need for expert or clinical intervention.

Among human listeners, recognition rates increased up to 191% (from 21.6% to 41.2%) relative to the original speech by using the module that corrects pronunciation errors. Several types of modified dysarthric speech signals are also supplied to a standard automatic speech recognition system. In that study, the proportion of words correctly recognized increased up to 121% (from 72.7% to 87.9%) relative to the original speech, across various parameterizations of the recognizer. This represents a significant advance towards human-to-human assistive communication software and human–computer interaction.

© 2012 Elsevier Ltd. All rights reserved.

*Keywords:* Speech transformation; Dysarthria; Intelligibility

## 1. Introduction

Speech communication can be fundamental to an individual's participation in society. However, this medium of expression is often disrupted by various types of physical disorder such as deafness, muteness, or reduced motor control. Dysarthria encapsulates various neuro-motor disorders that impair the physical production of speech. These impairments reduce the normal control of the primary vocal articulators but do not affect the abstract comprehension or production of meaningful, syntactically correct language. Reduced control of the vocal cords, for example, can result in aberrant voicing, and inadequate control of the velum may result in hypernasality. Lack of articulatory control can also lead 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 results in heavily slurred speech and a more diffuse and less differentiable vowel target space (Kent and Rosen, 2004).

<sup>☆</sup> This paper has been recommended for acceptance by Kathleen McCoy.

<sup>\*</sup> Tel.: +1 416 946 8573; fax: +1 416 946 8570.

*E-mail address:* [frank@cs.toronto.edu](mailto:frank@cs.toronto.edu)

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

Consider an individual with severe dysarthria who must travel into a city by public transportation. This activity might involve purchasing tickets, asking for directions, or indicating intentions to fellow passengers, all within noisy, hectic, and crowded environments. A personal portable communication device in this scenario (e.g., either held in the hand or attached to a wheelchair) would transform relatively unintelligible speech spoken into a microphone to make it more intelligible before being amplified and played over a set of speakers. This type of interaction easily generalizes to other scenarios in which an individual with a speech disability needs to communicate to those without significant experience interacting with this population.

Transforming one's speech in a way that preserves the natural prosody can also preserve extra-linguistic information, such as emotions, and is therefore a pertinent response to the limitations of current technology. This paper describes two experiments in which pitch-preserving transformations are used to make dysarthric speech signals more similar to speech from the general population. In the first experiment, the resulting signals are presented to human listeners to judge the intelligibility improvements possible in mediated human–human interaction. This work extends former work (Rudzicz, 2011a) for which a provisional US patent has been filed (US 78053/00002). In the second experiment, the transformed signals are presented to an automatic speech recognition system to judge the applicability of these techniques in human–computer interaction.

## 2. 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 (7.7 s vs 16.9 s), 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 et al., 2003), and command-and-control systems (Green et al., 2003) indicating that 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, 2011b). 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), although some research aims to rectify this (Barra-Chicote et al., 2010). 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 general communicative ability. Furthermore, the use of pre-recorded messages is implicitly based on the assumption that one can predict what one will wish to say at any given time. Even if the scenario or location is taken into account in specialized lists or vocabularies, the expressive output of such systems may still be needlessly limited.

### 2.1. Speech modification

Kain et al. (2007) proposed the voice transformation system shown in Fig. 1 which produced output signals by concatenating together original unvoiced segments with synthesized voiced segments that consisted of a superposition

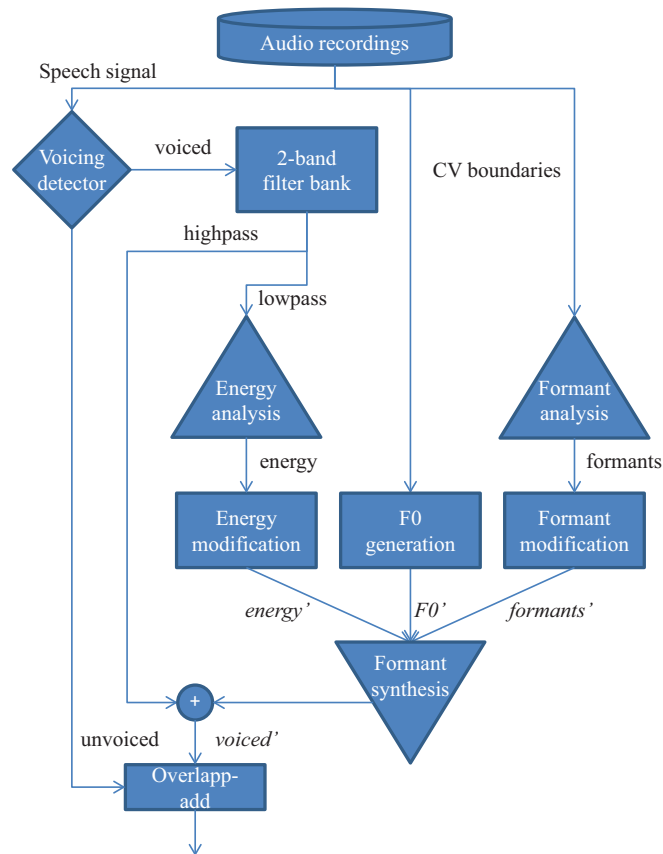


Fig. 1. Voice transformation system proposed by Kain et al. (2007).

of the original high-frequency band signal with synthesized low-frequency band formants. Although not explicitly explained in that paper, the high band may have been retained in order to preserve energy characteristics between 4 kHz and 8 kHz that are consistent with human speech (i.e., formant synthesis below 4 kHz alone would ignore such energy). Modifications to energy and formants were performed by Gaussian mixture mapping, as described below, in which learned relationships between dysarthric and non-dysarthric acoustics were used to produce output closer to the latter. Although Kain et al. (2007) performed extensive hand-tuning and manually identified formants in the input for the purposes of their study, in practice these processes can be automated to a large extent in real-time systems. 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 mispronunciations (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 24th order cepstral coefficients. 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 using aligned, phoneme-annotated, and orthographically identical sentences spoken by dysarthric and regular speakers. The probability distribution of the dysarthric source spectral features  $x$  was modeled using a Gaussian Mixture

Model (GMM) as the sum of  $D$  normal distributions with mean vector  $\mu$ , diagonal covariance matrix  $\Sigma$ , and prior probability  $\alpha$ :

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

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 dysarthric speech  $\hat{y}$ :

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

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.<sup>1</sup> 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 (68–87% accuracy) among a group of 18 naive human listeners each of whom annotated a total of 206 dysarthric test words (Hosom et al., 2003).

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. In that study, 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.

## 2.2. Automatic speech recognition for speakers with dysarthria

Applications of automatic speech recognition (ASR) for dysarthric speakers have been relatively pervasive, including automatic dictation of spontaneous text (Havstam et al., 2003), telephonic access to services, and the local control of machines (e.g., wheelchair, domestic appliances) (Hawley et al., 2007).

Early work in applying ASR to individuals with dysarthria almost exclusively involved the use of hidden Markov models (HMMs) whose parameters were trained to speech from the general population. Usually, these involved small-vocabulary recognition tasks with word-recognition rates significantly lower for dysarthric speakers. For example, given a vocabulary of 40 words, Rodman et al. (1985) reported mean word-recognition rates of 58.6% for dysarthric speakers compared with 95% for the general population. Modern approaches have closed this gap somewhat, but the disparity remains. Polur and Miller (2006), for example, produced ergodic HMMs that allow for ‘backwards’ state transitions. This ergodic structure is meant to capture aspects of dysarthric speech such as involuntary repetition and disruptions during sonorants (e.g., pauses) and revealed small improvements over the traditional baseline. Morales and Cox (2009) improved word-error rates by approximately 5% on severely dysarthric speech and approximately 3% on moderately dysarthric speech by building weighted transducers into an ASR system according to observed phonetic confusion matrices. The metamodels used in this work are very similar to those used by Matsumasa et al. (2009), except it also involved a language model, albeit one based on the highly restricted Nemours database (Menendez-Pidal et al., 1996). A commonality among all this work is that the actual articulatory behaviour of the dysarthric speech has not been taken into account.

Adapting HMM acoustic models trained to the general population given dysarthric data has also shown to improve accuracy, but not as clearly as training those models exclusively with dysarthric acoustics, especially in the more

<sup>1</sup> This model can also be used to measure the difference between any two types of speech, as is often the case for non-pathological voices (Toda et al., 2005).

severe cases (Raghavendra et al., 2001; Sanders et al., 2002). There is some controversy on this matter, however, with some recent studies suggesting that speakers with especially low intelligibility obtain better rates of recognition from maximum *a posteriori* adaptive models (Sharma and Hasegawa-Johnson, 2010).

### 3. The TORGO-based transformations

We propose a serialized sequence of acoustic transformations of signals uttered by speakers with dysarthria. Each of these transformations is designed in response to a unique effect of dysarthria on intelligibility as determined by observations described by Rudzicz et al. (2008). 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 6.

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 described in Section 4 that synthesizes phonemes using a standard method of linear predictive coding with a pronunciation lexicon and part-of-speech tagger, and selection of intonation parameters (Taylor et al., 1998). This system is invoked by providing the expected text uttered by the dysarthric 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 our transformations, which are outlined in Fig. 2. These components allow for a cascade of one transformation followed by another, although we can also perform these steps independently. In all cases, the spectrogram is derived with the fast Fourier transform given 2048 bins on the range of 0–5 kHz in which all formants in which we are interested occur. 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 fundamental frequency ( $F_0$ ) 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.

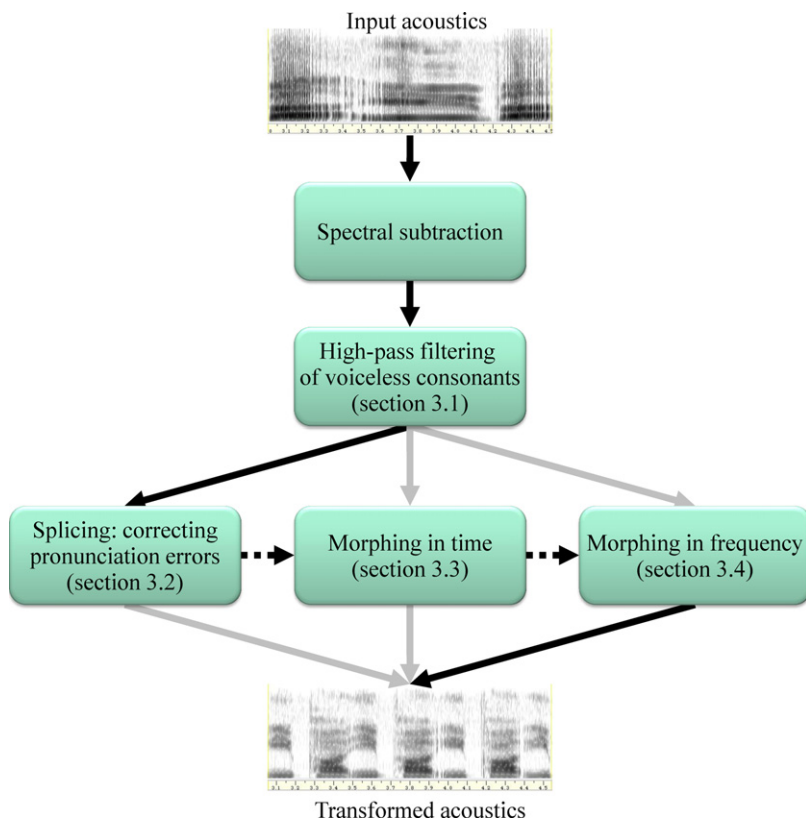


Fig. 2. Outline of the proposed transformation system. The black path indicates the cascade to be used in practice. Solid arrows indicate paths taken during evaluation.

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 10th-order high-pass Butterworth filter converted to the discrete equivalent through the impulse-invariant discretization method implemented by the difference equation

$$y[n] = \sum_{k=1}^{10} a_k y[n-k] + \sum_{k=0}^{10} b_k x[n-k]. \quad (3)$$

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 is 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).

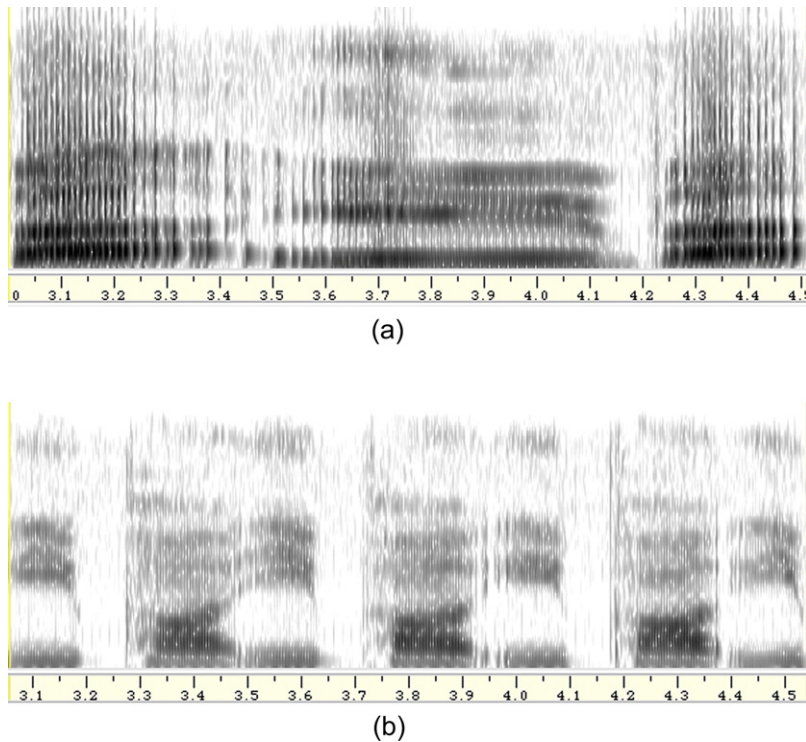


Fig. 3. Repetitions of */iy p ah/* over 1.5 s by (a) a male speaker with athetoid CP, and (b) a female control in the TORGO database. Dysarthric speech is notably slower and more strained than regular speech.

**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. Tempo morphing

Fig. 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 et al., 2012b). 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 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  $\theta$  is predicted by

$$\theta_k^{(F)} = \theta_j^{(F)} + 2\pi F(j - k). \quad (4)$$

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. Frequency morphing

Formant trajectories inform the listener as to the identities of vowels, but the vowel space of dysarthric speakers tends to be constrained (Kent et al., 1999). 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 14th-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,<sup>2</sup> formant candidates are identified at each frame 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. An example of the results of this morphing technique is shown in Fig. 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. Experiment 1: intelligibility with human listeners

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). For example, transcription of semantically unpredictable sentences, the diagnostic rhyme test, the diagnostic medial consonant test, and random phonetic foils are all frequently used, typically according to the material. Since our data include full sentence prompts from the TIMIT database, single-word tests of intelligibility were not applicable. Hustad (2006) suggested that orthographic transcriptions provide a more accurate predictor of intelligibility of dysarthric speech than the more subjective estimates used in clinical settings, e.g., Enderby (1983). That study had 80 listeners who transcribed audio and showed that intelligibility (as measured by the proportion of correct words identified in transcription according to Yorkston and Beukelman (1981)) 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.

<sup>2</sup> Accidentally inserted vowels are also included here, unless previously removed by the splicing technique in Section 3.2.



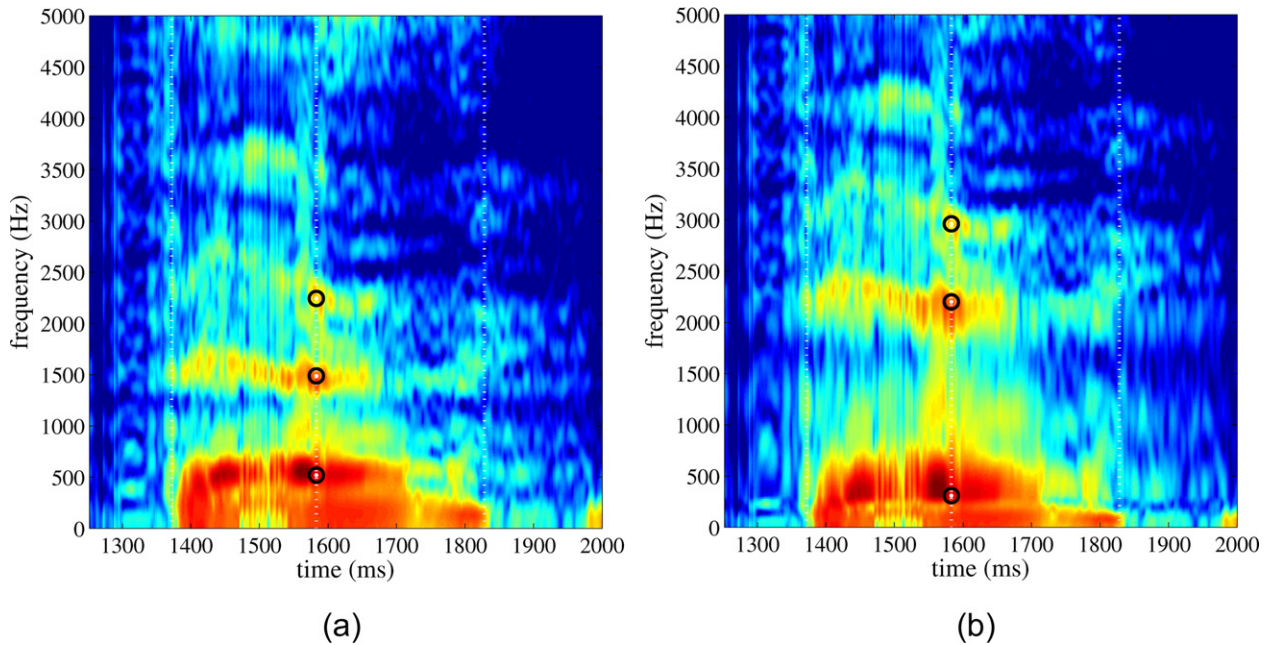


Fig. 4. Spectrograms for (a) the dysarthric original and (b) the frequency-modified renditions of the word *fear*. Circles represent indicative formant locations.

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 produced 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 about 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 (without replacement) 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.

TORGO is a database of aligned acoustics and articulographic data that includes data from eight individuals with dysarthria caused by cerebral palsy or amyotrophic lateral sclerosis and age- and gender-matched non-dysarthric subjects (Rudzicz et al., 2012b). Each of the individuals with speech impediments are associated with a Frenchay assessment of speech-motor function (Enderby, 1983) administered by a speech-language pathologist. The average dysarthric speaker in this database as determined by an average of individual scores would be classified as having ‘moderate’ dysarthria, although several speakers are classified as ‘severely’ dysarthric, though verbal. Data were recorded according to the total commitment of time (i.e., 3 h for each individual), resulting in an average of 415.4 utterances from each dysarthric individual (approximately 60% of which are single-word prompts and 40% of which are whole sentences). These data have been de-identified and made anonymous – it is not possible to identify the individuals based on the available data. Acoustic data were obtained by one head-mounted and one multi-channel directional microphone. Articulatory data were obtained by electromagnetic articulography, which allows the measurement of the tongue and other articulators during speech, and by 3D reconstruction from binocular video sequences. The stimuli

Table 1

Percentage of *words* correctly identified by each listener (L0\*) relative to the expected sequence. Sections 3.2–3.4 discuss the ‘Splice’, ‘Tempo’, and ‘Freq.’ techniques, respectively.

	Word accuracy (%)						Phoneme accuracy (%)					
	Orig.	GMM	Synth.	Spl.	Tmp.	Freq.	Orig.	GMM	Synth.	Spl.	Tmp.	Freq.
L01	22.1	15.6	82.0	40.2	34.7	35.2	52.0	43.1	98.2	64.7	47.8	55.1
L02	27.8	12.2	75.5	44.9	39.4	33.8	57.8	38.2	92.9	68.9	50.6	53.3
L03	38.3	14.8	76.3	37.5	12.9	21.4	50.1	41.4	96.8	57.1	30.7	46.7
L04	24.7	10.8	72.1	32.6	22.2	18.4	51.6	33.8	88.7	51.9	43.2	45.0
L05	21.6	8.2	70.7	41.2	18.4	23.8	49.3	35.4	90.8	67.0	41.3	46.1
L06	28.1	16.3	74.0	39.8	24.6	32.8	54.9	39.2	84.3	65.2	38.7	43.4
L07	29.2	15.3	76.2	41.1	26.3	28.6	53.8	40.1	93.5	63.6	43.9	51.8
L08	26.7	11.4	75.6	39.0	24.9	27.3	50.7	38.0	91.2	62.1	41.3	47.2
Avg.	27.3	13.1	75.3	39.5	25.4	27.7	52.5	38.7	92.1	62.6	42.2	48.6

were obtained from a variety of sources including the TIMIT database (Zue et al., 1989), lists of identified phonetic contrasts (Kent et al., 1989), and assessments of speech intelligibility (e.g., Yorkston and Beukelman (1981)). This database is made available by the Linguistic Data Consortium (Rudzicz et al., 2012a).

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’ and its default parameters. This system is based on Festival which can use linguistic pre-processing (e.g., part-of-speech tagging) and decision trees (e.g., letter-to-sound rules) (Taylor et al., 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 et al. (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 24th-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 3.2–3.4 above. Table 1 shows 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.

When comparing the three transformations individually (i.e., not in combination), only the splicing technique gave statistically significant improvements in word intelligibility over the original speech (paired  $t(7) = -5.27$ ,  $p < 0.005$ ). However, all three methods gave statistically significant improvements in phoneme intelligibility over the original speech; i.e.,  $t(7) = -5.71$ ,  $p < 0.001$  for splicing,  $t(7) = 5.86$ ,  $p < 0.001$  for temporal modifications, and  $t(7) = 2.71$ ,  $p < 0.05$  for frequency modifications. It is clear that the purely synthetic speech has a far greater intelligibility than other approaches, more than doubling the average accuracy of our proposed 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 (which is consistent with results with semantically unpredictable sentences from the Blizzard challenge (Chen et al., 2011)) is not an argument against the type of modifications proposed here and by Kain et al. (2007), since our aim

Table 2

Average ratings of naturalness on the 6-point scale from 1 (very natural) to 6 (very unnatural) according to the original speech, the GMM modified speech, purely synthesized speech, and the three TORGO modifications.

	Orig.	GMM	Synth.	Spl.	Tmp.	Freq.
L05	3.15	5.91	2.59	2.04	2.87	3.45
L06	2.18	5.84	3.08	2.42	2.63	4.09
L07	2.87	5.72	2.68	1.19	2.41	4.24
L08	2.56	5.91	2.33	1.4	2.5	3.72
Avg.	2.69	5.84	2.92	1.51	2.6	3.88

is to avoid the use of impersonal and invariant utterances. Further uses of acoustic modifications not attainable by text-to-speech synthesis are discussed in Section 6.

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. We compare the Frenchay scores of intelligibility with the acoustic energy of sonorants, as computed over 8 ms Hamming windows. Over all data, we find Pearson correlation coefficients of  $-0.34$ ,  $-0.66$ , and  $-0.81$  between the acoustic energy and the word, sentence, and conversational intelligibility scores, respectively. Tangentially, Neel (2009) found that loud or amplified speech from individuals with Parkinson's disease was more intelligible to human listeners than quieter speech, although the relevance of that work here may be mitigated by other differences between Parkinson's disease and cerebral palsy.

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 and 48% correct dysarthric vowel recognition (compared with 52.9% in Table 1)). 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).

A simple Likert-type scale was presented to each of four participants in order to measure the naturalness of the synthesized and transformed speech. Subjects were asked to rate naturalness as defined as “sounding like everyday common speech, allowing for the possibility that the speaker has a speaking disability”. Specifically, each utterance is rated according to the 6-point scale used by Kraft and Portele (1995) where a score of 1 is ‘very natural’ and a score of 6 is ‘very unnatural’. These results are averaged according to the sources, approximating a continuous scale, as provided in Table 2. The original dysarthric speech is rated between ‘very natural’ and ‘natural’. The splicing technique results in speech that sounds significantly more natural than the synthesized speech (paired  $t(3) = 7.41$ ,  $p < 0.01$ ). Somewhat surprisingly, the benefit of modifying the tempo of the signal does not result in significant improvements. While this test provides an evaluation of naturalness, it does not measure whether extra-lexical content such as specific emotional or lexical content is retained.

It is also important that our synthesized or transformed speech resemble the original speaker in subjective perceptual terms. To test the similarity of our transformations, we follow the method of Wester et al. (2010) by presenting pairs of utterances to human listeners and asking the binary question, “do these utterances sound like they came from the same speaker?”. A negative response is scored as a zero and a positive response is scored as a one. Wester et al. (2010) suggest that a more fine-grained Likert-type scale may be ‘too difficult for listeners.’ In each case, the original dysarthric utterance was presented along with its synthesized or modified equivalent (i.e., the lexical content was identical). Table 3 shows the average result of the binary similarity question across all non-original utterances for each listener. Each of the TORGO-based transformations sound significantly like the original speaker, as measured relative to random chance (i.e., scores of 0.5); specifically, paired one-tailed tests of  $t(3) = 33.64$ ,  $p < 0.0001$  for splicing,  $t(3) = 21.82$ ,  $p < 0.0001$  for temporal modifications, and  $t(3) = 10.18$ ,  $p < 0.005$  for frequency modifications. Both the GMM and synthetic methods sound significantly unlike the original speaker, as measured relative to random chance (i.e., scores of 0.5); specifically, paired one-tailed tests of  $t(3) = -7.85$ ,  $p < 0.005$  for the GMM method and  $t(3) = -\infty$ ,

Table 3

Average similarity scores between the original dysarthric speech and its synthesized or modified equivalents. A score of 0 indicates dissimilarity and a score of 1 indicates similarity.

	GMM	Synth.	Spl.	Tmp.	Freq.
L05	0.19	0.0	0.92	0.78	0.67
L06	0.28	0.0	0.95	0.81	0.73
L07	0.33	0.0	0.98	0.85	0.77
L08	0.22	0.0	0.93	0.82	0.70
Avg.	0.26	0.0	0.95	0.82	0.72

$p = 0$  for the synthetic method. This validates a key goal of this work, specifically to produce voices perceptually similar to the original speaker. Interestingly, no single comparison between the original and completely synthesized speech suggested similarity.

## 5. Experiment 2: accuracy with ASR

Despite recent advances in automatic speech recognition for dysarthric speech (Rudzicz, 2011b), recognition accuracy is still limited by the relatively unconstrained nature of modern underlying statistical methods and short-time observation windows. Several fundamental phenomena of dysarthria such as increased disfluency, longer sonorants, and imprecise vowels (Rudzicz et al., 2008) are not readily encoded in popular statistical frameworks such as hidden Markov models (HMMs) or dynamic Bayes networks (DBNs). However, some of these phenomena may be offset by our proposed transformations prior to the use of ASR. In this experiment we transform the same dysarthric speech data as in the previous experiment using each of the proposed transformations, as well as their union, and present these to an HMM-based classifier. These modifications are independent, so their order of execution is irrelevant. As before, we compare the original dysarthric speech, the dysarthric speech transformed by Gaussian mixture model, and the purely synthetic speech with the results of each of the proposed transformations. In each case, we evaluate utterances using traditional HMMs trained with large amounts of data from the general population (i.e., the complete Wall Street Journal (WSJ0 and WSJ1) acoustic corpus) in order to simulate a typical listener. That is, we do not use any data from dysarthric speakers in training (or adaptation) in order to obtain models representing the general population. Each phoneme is modeled by left-to-right, tri-state monophone HMMs and a single-state short-pause model. In each case, monophones are strung together into word networks according to the CMU pronunciation dictionary. We vary the number of Gaussians per HMM state from 1 to 32 in powers of 2.

Table 4 shows the percentage of words correctly identified across the various acoustic conditions and number of Gaussians per state. Clearly, improvements are possible if we pre-process the evaluation data using the proposed methods. All transformation methods (individually and in combination) result in significantly better accuracy than the original speech at the 95% level of confidence, given paired  $t$ -tests. Unlike intelligibility tests with human listeners, the frequency warping was more suited to automatic recognition than the other two modifications individually, but the highest rates of recognition were uniformly obtained using all techniques with relative reductions in error of between 51.6% and 55.7% across the specificity of the state observation models. Interestingly, combining the splicing and

Table 4

Percentage of words correctly identified by hidden Markov model recognition systems varying by number of Gaussians per state across various acoustic conditions.

No. Gaussians	Orig.	GMM	Synth.	Splice	Tempo	Freq.	Splice + Tempo	Splice + Freq.	Tempo + Freq.	All
1	72.2	51.9	93.6	85.3	82.1	87.0	83.9	87.3	86.9	87.2
2	72.7	52.2	93.6	85.5	82.5	87.4	84.0	87.7	87.5	87.9
4	74.4	52.8	94.3	85.7	82.8	87.3	84.1	87.7	87.6	88.3
8	75.6	53.1	94.5	85.8	83.6	87.9	84.8	88.4	88.1	88.7
16	77.1	53.8	94.9	86.0	84.2	88.2	85.0	88.6	88.4	89.1
32	77.7	54.0	95.2	86.1	84.4	88.2	85.3	88.7	88.4	89.2

timing techniques without frequency scaling resulted in worse accuracy than the splicing technique alone, whereas all other combinations had better accuracy than their component parts alone.

## 6. Discussion

This work represents a step towards speech modification systems for human–human and human–computer interaction. The results in Section 4 shows 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.

Similarly, controlling computers and other devices with one's voice is a desirable alternative to physical methods of computer interaction for individuals with speech disabilities (Hosom et al., 2003). Rather than exploring alternative speech models or classification algorithms to cope with the effects of dysarthria, our approach is to use unmodified traditional ASR tuned to the general population with signals that are modified to better match that target. This mimics situations in which only speaker-independent models for the general population are available, e.g., in typical telephonic applications. Our results show relative error reductions in word recognition of an order of magnitude greater than previous work based on alternative models (i.e., up to 55.7% compared with 5% Morales and Cox, 2009); however, these results are not directly comparable due to the differing nature of the data. Future work should determine how each transformation component effects ASR accuracy. For instance, frequency-based transformations may result in acoustic spaces with lower entropy (which is therefore easier to model statistically) since these modifications are directed towards a single canonical standard voice. Also, the removal of unintended repetition may directly influence the prevalence of insertion errors. We are interested in studying how the use of these modified speech signals may be used to adapt existing models for the general population. Specifically, we are interested in applying standard maximum likelihood linear regression on speaker-independent models given only transformed speech signals; the result may be better tuned to the specific characteristics of these adjustments.

A prototypical client-based application based on our research for unrestricted speech transformation of novel sentences to be used in practice is currently in development. Such work will involve improving factors such as accuracy and accessibility for individuals whose neuro-motor 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. 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 et al. (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 et al., 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. Since these traditional models often construct words out of identified phonemes, out-of-vocabulary words would merely be represented by their identified phoneme sequences and would not require additional attention.

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.

## Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.csl.2012.11.001>.

## References

- Allen, J., Hunnicutt, M.S., Klatt, D.H., Armstrong, R.C., Pisoni, D.B., 1987. *From Text to Speech: The MITalk System*. Cambridge University Press, New York, NY, USA.
- Arbisi-Kelm, T., 2010. Intonation structure and disfluency detection in stuttering. *Laboratory Phonology* 10 (4), 405–432.
- Banno, H., Hata, H., Morise, M., Takahashi, T., Irino, T., Kawahara, H., 2007. Implementation of realtime STRAIGHT speech manipulation system: report on its first implementation. *Acoustical Science and Technology* 28 (3), 140–146.
- Barra-Chicote, R., Yamagishi, J., King, S., Montero, J.M., Macias-Guarasa, J., 2010 May. Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. *Speech Communication* 52 (5), 394–404.
- Black, A.W., Lenzo, K.A., 2007. Building synthetic voices. <http://www.festvox.org/festvox/bsv.ps.gz>
- Chen, L.-H., Yang, C.-Y., Ling, Z.-H., Jiang, Y., Dai, L.-R., Hu, Y., Wang, R.-H., 2011. The usc system for blizzard challenge 2011. [http://www.festvox.org/blizzard/bc2011/USTC\\_Blizzard2011.pdf](http://www.festvox.org/blizzard/bc2011/USTC_Blizzard2011.pdf)
- Czyzewski, A., Kaczmarek, A., Kostek, B., 2003. Intelligent processing of stuttered speech. *Journal of Intelligent Information Systems* 21 (2), 143–171.
- Enderby, P.M., 1983. *Frenchay Dysarthria Assessment*. College Hill Press, London, UK.
- Green, P., Carmichael, J., Hatzis, A., Enderby, P., Hawley, M., Parker, M., 2003. Automatic speech recognition with sparse training data for dysarthric speakers. In: *Proceedings of Eurospeech 2003*, Geneva, pp. 1189–1192.
- Hasegawa-Johnson, M., Fleck, M., 2007. International Speech Lexicon Project. <http://www.isle.illinois.edu/dict/>
- Havstam, C., Buchholz, M., Hartelius, L., 2003 August. 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).
- Hawley, M.S., Enderby, P., Green, P., Cunningham, S., Brownsell, S., Carmichael, J., Parker, M., Hatzis, A., O'Neill, P., Palmer, R., 2007 June. A speech-controlled environmental control system for people with severe dysarthria. *Medical Engineering & Physics* 29 (5), 586–593.
- Hayes, M.H., 1999. *Digital Signal Processing. Schaum's Outlines*, McGraw Hill.
- Hess, W.J., 2008. Pitch and voicing determination of speech with an extension toward music signal. In: Benesty, J., Sondhi, M.M., Huang, Y. (Eds.), *Speech Processing*. Springer, Berlin, Germany.
- Hosom, J.-P., Kain, A.B., Mishra, T., van Santen, J.P.H., Fried-Oken, M., Staehely, J., 2003 April. Intelligibility of modifications to dysarthric speech. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 1, pp. 924–927.
- Hustad, K.C., 2006. Estimating the intelligibility of speakers with dysarthria. *Folia Phoniatrica et Logopaedica* 58 (3), 217–228.
- Hustad, K.C., Beukelman, D.R., 2002 June. Listener comprehension of severely dysarthric speech: effects of linguistic cues and stimulus cohesion. *Journal of Speech Language and Hearing Research* 45, 545–558.
- Hux, K., Rankin-Erickson, J., Manasse, N., Lauritzen, E., January 2000. Accuracy of three speech recognition systems: case study of dysarthric speech. *Augmentative and Alternative Communication (AAC)* 16 (3), 186–196. URL <http://dx.doi.org/10.1080/07434610012331279044>
- Kain, A.B., Hosom, J.-P., Niu, X., van Santen, J.P., Fried-Oken, M., Staehely, J., 2007 September. Improving the intelligibility of dysarthric speech. *Speech Communication* 49 (9), 743–759.
- Kawahara, H., 2006. STRAIGHT, exploitation of the other aspect of VOCODER: perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology* 27 (6), 349–353.
- Kawahara, H., de Cheveigné, A., Banno, H., Takahashi, T., Irino, T., 2005 September. Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT. In: *Proceedings of INTERSPEECH 2005*, pp. 537–540, [http://www.isca-speech.org/archive/interspeech\\_2005/i05\\_0537.html](http://www.isca-speech.org/archive/interspeech_2005/i05_0537.html)
- Kawahara, H., Matsui, H., April 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. Vol. 1, I-256–I-259 vol. 1.
- Kent, R.D., Rosen, K., 2004. Motor control perspectives on motor speech disorders. In: Maassen, B., Kent, R., Peters, H., Lieshout, P.V., Hulstijn, W. (Eds.), *Speech Motor Control in Normal and Disordered Speech*. Oxford University Press, Oxford, pp. 285–311, Ch. 12.
- Kent, R.D., Weismer, G., Kent, J.F., Rosenbek, J.C., 1989. Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders* 54, 482–499.
- Kent, R.D., Weismer, G., Kent, J.F., Vorperian, H.K., Duffy, J.R., 1999. Acoustic studies of dysarthric speech: methods, progress, and potential. *Journal of Communication Disorders* 32 (3), 141–186.
- Kida, Y., Kawahara, T., 2005. Voice activity detection based on optimally weighted combination of multiple features. In: *Proceedings of INTERSPEECH-2005*, pp. 2621–2624.
- Kraft, V., Portele, T., 1995. Quality evaluation of five speech synthesis systems. *Acta Acoustica* 3, 351–365.
- Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory* 10 (8), 707–710.
- Matsumasa, H., Takiguchi, T., Arika, Y., Li, I.-C., Nakabayashi, T., 2009 August. Integration of metamodel and acoustic model for dysarthric speech recognition. *Journal of Multimedia* 4 (4), 254–261.
- Menendez-Pidal, X., Polikoff, J.B., Peters, S.M., Leonzjo, J.E., Bunnell, H., 1996 October. The nemours database of dysarthric speech. In: *Proceedings of the Fourth International Conference on Spoken Language Processing*, Philadelphia, PA, USA.
- Messina, J.J., Messina, C.M., April 2007. Description of AAC devices. <http://www.coping.org/specialneeds/assistechn/aacdev.htm>
- Morales, S.O.C., Cox, S.J., 2009. Modelling errors in automatic speech recognition for dysarthric speakers. *EURASIP Journal on Advances in Signal Processing*.
- Moulines, E., Charpentier, F., 1990 December. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication* 9, 453–467, URL <http://portal.acm.org/citation.cfm?id=116058.116064>

- Nakatani, C., 1993. A speech-first model for repair detection and correction. In: *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 46–53.
- Neel, A.T., 2009 August. Effects of loud and amplified speech on sentence and word intelligibility in parkinson disease. *Journal of Speech Language and Hearing Research* 52, 1021–1033.
- O’Shaughnessy, D., 2000. *Speech Communications—Human and Machine*. IEEE Press, New York, NY, USA.
- O’Shaughnessy, D., 2008. Formant estimation and tracking. In: Benesty, J., Sondhi, M.M., Huang, Y. (Eds.), *Speech Processing*. Springer, Berlin, Germany.
- Patel, R., December 1998. Control of prosodic parameters by an individual with severe dysarthria. Tech. rep., University of Toronto. URL [http://vismod.media.mit.edu/pub/masters\\_paper.doc](http://vismod.media.mit.edu/pub/masters_paper.doc)
- Polur, P.D., Miller, G.E., 2006 October. Investigation of an HMM/ANN hybrid structure in pattern recognition application using cepstral analysis of dysarthric (distorted) speech signals. *Medical Engineering and Physics* 28 (8), 741–748.
- Portnoff, M.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, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1992. *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge University Press.
- Raghavendra, P., Rosengren, E., Hunnicutt, S., 2001 December. An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems. *Augmentative and Alternative Communication (AAC)* 17 (4), 265–275.
- Rodman, R., Moody, T., Price, J., 1985. Speech recognizer performance with dysarthric speakers: a comparison of two training procedures. *Speech Technology* 1, 65–71.
- Rosen, K., Yampolsky, S., 2000 Jan. Automatic speech recognition and a review of its functioning with dysarthric speech. *Augmentative & Alternative Communication* 16 (1), 48–60, URL <http://dx.doi.org/10.1080/07434610012331278904>
- Rudzicz, F., 2011a. Acoustic transformations to improve the intelligibility of dysarthric speech. In: *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (SLPAT2011) at the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP2011)*, Edinburgh Scotland, July.
- Rudzicz, F., 2011b. Production knowledge in the recognition of dysarthric speech. Ph.D. Thesis. University of Toronto, Department of Computer Science.
- Rudzicz, F., Hirst, G., van Lieshout, P., Penn, G., Shein, F., Namasivayam, A., Wolff, T., 2012a. TORGO Database of Dysarthric Articulation. <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2012S02>
- Rudzicz, F., Namasivayam, A.K., Wolff, T., 2012b. The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation* 46 (4), 523–541.
- Rudzicz, F., van Lieshout, P., Hirst, G., Penn, G., Shein, F., Wolff, T., 2008 December. Towards a comparative database of dysarthric articulation. In: *Proceedings of the eighth International Seminar on Speech Production (ISSP’08)*, Strasbourg, France.
- Sanders, E., Ruiters, M., Beijer, L., Strik, H., 2002 September. Automatic Recognition of Dutch dysarthric speech: a pilot study. In: *7th International Conference on Spoken Language Processing*, Denver, Colorado.
- Schroeter, J., 2008. Basic principles of speech synthesis. In: Benesty, J., Sondhi, M.M., Huang, Y. (Eds.), *Speech Processing*. Springer, Berlin, Germany.
- Sethares, W.A., 2007. *Rhythm and Transforms*. Springer, Berlin, Germany.
- Sharma, H.V., Hasegawa-Johnson, M., 2010. State-transition interpolation and MAP adaptation for HMM-based dysarthric speech recognition. In: *Proceedings of the NAACL HLT Workshop on Speech and Language Processing for Assistive Technologies*, pp. 72–79.
- Snell, R.C., Milinazzo, F., 1993. Formant location from LPC analysis data. *IEEE Transactions on Speech and Audio Processing* 1 (April (2)).
- Spiegel, M.F., Altom, M.J., Macchi, M.J., Wallace, K.L., 1990. Comprehensive assessment of the telephone intelligibility of synthesized and natural speech. *Speech Communication* 9 (4), 279–291, URL <http://www.sciencedirect.com/science/article/B6V1C-48V21K0-FN/2/724c382576c44981bcc8e8bba88626bd>
- Stevens, K.N., 1998. *Acoustic Phonetics*. MIT Press, Cambridge, Massachusetts.
- Stylianou, Y., 2008. Voice transformation. In: Benesty, J., Sondhi, M.M., Huang, Y. (Eds.), *Speech Processing*. Springer, Berlin, Germany.
- Taylor, P., Black, A.W., Caley, R., 1998. The architecture of the Festival speech synthesis system. In: *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*. Jenolan Caves, Australia, pp. 147–151.
- Toda, T., Black, A.W., Tokuda, K., 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, H., Torgoman, A.S.E., 2009. Towards the improvement of automatic recognition of dysarthric speech. In: *International Conference on Computer Science and Information Technology*, IEEE Computer Society, Los Alamitos, CA, USA, pp. 277–281.
- Toth, A.R., Black, A.W., 2005. Cross-speaker articulatory position data for phonetic feature prediction. In: *Proceedings of Interspeech 2005*, Lisbon, Portugal.
- Wester, M., Dines, J., Gibson, M., Liang, H., Wu, Y.-J., Saheer, L., King, S., Oura, K., Garner, P.N., Byrne, W., Guan, Y., Hirsimäki, T., Karhila, R., Kurimo, M., Shannon, M., Shiota, S., Tian, J., Tokuda, K., Yamagishi, J., 2010 September. Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project. In: *Proc. of 7th ISCA Speech Synthesis Workshop*, Kyoto, Japan.
- Yan, Q., Vaseghi, S., Zavarehei, E., Milner, B., Darch, J., White, P., Andrianakis, I., 2007. Formant tracking linear prediction model using HMMs and Kalman filters for noisy speech processing. *Computer Speech and Language* 21, 543–561.
- Yorkston, K.M., Beukelman, D.R., 1981. *Assessment of Intelligibility of Dysarthric Speech*. C.C. Publications Inc., Tigard, Oregon.
- Zue, V., Seneff, S., Glass, J., 1989. Speech Database Development: TIMIT and Beyond. In: *Proceedings of ESCA Tutorial and Research Workshop on Speech Input/Output Assessment and Speech Databases (SIOA-1989)*, Vol. 2. Noordwijkerhout, The Netherlands, pp. 35–40.