FORMANT RE-SYNTHESIS OF DYSARTHRIC SPEECH

Alexander Kain, Xiaochuan Niu, John-Paul Hosom, Qi Miao, Jan van Santen

Center for Spoken Language Understanding
OGI School of Science & Engineering
Oregon Health & Science University, Portland, Oregon, USA
{kain,xiaochua,hosom,miaoqi,vansanten}@cslu.ogi.edu

ABSTRACT

Dysarthria is a motor speech disorder that is often associated with irregular phonation (e.g. vocal fry) and amplitude, incoordination of articulators, and restricted movement of articulators, among other problems. The present study is part of a project on voice transformation systems for dysarthria, with the goal of producing intelligibility-enhanced speech. We report on a procedure in which formants and energies are estimated from dysarthric speech; next, these trajectories are modified to more closely approximate desired targets; finally, transformed speech is generated using formant synthesis. Results indicate that the transformation step enhances intelligibility, and that removal of vocal fry enhances perceived quality. However, the initial step of stylizing the formant trajectories results in a decrement in intelligibility, thereby reducing the net impact of the process.

1. INTRODUCTION

Dysarthria is a motor speech disorder due to weakness or poor coordination of the speech muscles. Affected muscles include the lungs, larynx, oro- and nasopharynx, soft palate, and articulators (lips, tongue, teeth, and jaw). The degree to which these muscle groups are compromised determines the particular pattern of speech impairment. For example, poor lung function affects the overall volume or loudness, while problems with specific articulators may cause mispronunciations of certain phonemes. There is a great variety of diseases that can cause dysarthria, including Parkinson's, Multiple Sclerosis, and strokes.

We are currently investigating algorithms that, when implemented on a wearable device, can enable people with dysarthria to be understood by the general population. Although there are already devices on the market [1] that address certain forms of dysarthria, these devices, not substantially different from fine-tuned filters and amplifiers, do not take advantage of recent advances in speech technology and are therefore limited in the assistance they provide. While there is no doubt that such devices can help certain types of dysarthria, many dysarthric persons suffer from speech problems that require forms of speech modification that are much more complex. Among these problems are:

- Irregular sub-glottal pressure, resulting in distracting loudness variation
- Absence or poor control of voicing
- Systematic or variable mispronunciations of certain phoneme groups, resulting in certain sounds becoming indistinguishable or unrecognizable

- Prosodic problems, due to problems in pitch control

For these difficult problems, new approaches are needed that analyze the speech signal at the acoustic, articulatory, phonetic, and linguistic levels. As a first step toward such an approach, we propose a transformation system whose core idea is to extract acoustic parameters from the input speech signal that are especially relevant to speech intelligibility, modify those parameters, and then synthesize a new speech signal from them. It is likely that any specific form of the proposed system will only work for a subgroup of dysarthrias, due to the widely different forms of the disorder.

2. METHOD

2.1. Overview

Our overall strategy is to improve intelligibility by the manipulation of a small set of highly relevant speech features. We are motivated by the observation that formant synthesizers, which control only a very small set of speech features, may not sound very natural but are highly intelligible. Therefore, our selection of speech features is very similar to those used in formant synthesizers.

Figure 1 shows a flowchart of the system used in our approach. The dysarthric input speech waveform is subjected to several analyses that extract energy, F_0 and formant information. For voiced regions, these features are subsequently modified or, as is the case for F_0, completely re-generated. The modified features are then used as input to a formant synthesizer. Finally, the original unvoiced regions and the modified voiced regions are assembled into an output speech waveform.

Note that the system does not make any changes to phoneme durations. Also, no attempts are made to use additional, more conventional modifications such as compression/expansion of the speech dynamics or amplifications of certain frequency regions.

2.2. Speech Data

In order to deal with the formidable difficulty of the task, we limited ourselves to studying consonant-vowel-consonant (CVC) contexts from a special-purpose database. The speech data used for training, transformation and evaluation were utterances of one dysarthric speaker and one non-dysarthric (normal) speaker. Each speaker read 278 isolated, monosyllabic, nonsense CVC "words". The speech signal was recorded and stored in 16k Hz, 16-bit PCM format.

The vowels in the CVC words consisted of 4 front vowels (/i:/, /I/, /E/, and /@/; see [2] for pronunciation) and 4 back vowels (/A/, /^/, /U/, and /u/). These vowels are representative of typical vocal-tract configurations in American English. The consonants consisted of 6 stops (/p/, /b/, /t/, /d/, /k/, and /g/), 4 fricatives (/v/, /s/, /z/, and /S/), and 3 approximants (/l/, /j/, and /w/), covering
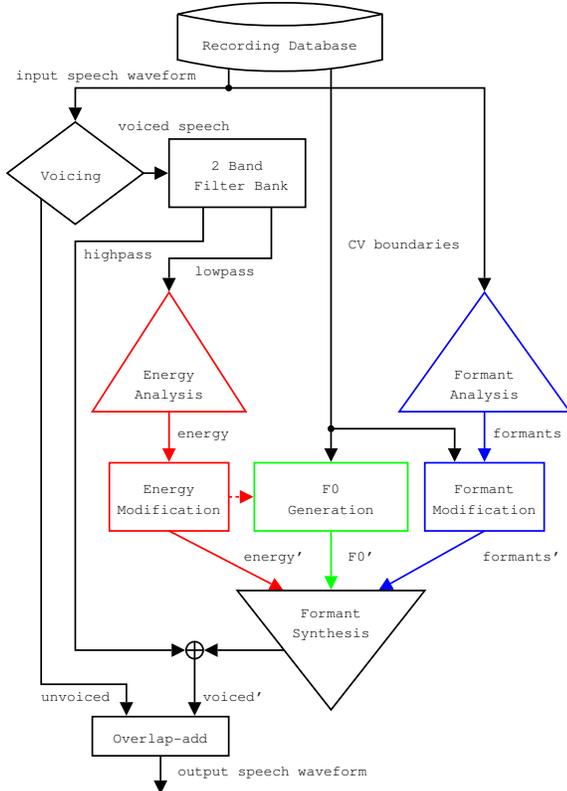
Fig. 1. Flowchart of the modification system.

most places of articulation and voicing distinctions among American English consonants. We manually segmented each utterance into a sequence of phoneme labels with time alignments.

The complete set of recordings was divided into two groups, A and B. Utterances from Group A were used for training and evaluating intelligibility, and utterances from Group B were used for evaluating speech quality. Group A was divided into two non-overlapping partitions, one of 228 utterances used for training the formant mapping function, and one of 50 utterances used for the intelligibility evaluation. The number of occurrences of each vowel in the intelligibility evaluation was 6 or 7. Group B consisted of 50 utterances from Group A's training partition, and was used for the evaluation of speech quality.

2.3. Analysis

2.3.1. Feature extraction

We used the ESPS waves+ software package [3] to extract voicing, F_0 and formant information. Additionally, we converted the phonetic labeling into consonant-vowel boundary information; while a production system would need to estimate these boundaries from the speech signal, we use the manual segmentation information directly. Voiced regions of speech are divided into a low band (0–4 kHz) and a high band (4–8 kHz). We calculate the energy of the lower band using root mean squared values multiplied by an asymmetric trapezoidal window (this window matches the final synthesis window).

2.3.2. Formant targets

It is well known that the formant pattern of a vowel is an important acoustical correlate of vowel identity. The first formant (F_1) is inversely related to the perception of vowel height, and the difference between the first (F_1) and the second (F_2) formant is related to the perception of vowel backness. Formant analysis may help us to understand the acoustical and articulatory reasons that impair the intelligibility of vowel sounds in dysarthric speech. By comparing the formant pattern of vowels in dysarthric speech with that in normal speech, we also expect to find a way to improve the intelligibility of dysarthric speech. In the present study, F_1 and F_2 modifications were considered. This is partly because F_1 and F_2 are perceptually more important than other formants, and partly because the formant tracker we used could provide more reliable F_1 and F_2 trajectories than other formant trajectories.

It has been assumed that there exists a target vocal-tract configuration during the production of each monophthong, and that this configuration corresponds to a certain formant pattern, which can be measured from the acoustic data at a stable point or section of the vowel that is least influenced by context. There have been different ways of choosing the sampling point or section in the studies of the formant characteristics of vowels. Stevens and House [4] studied formant values at temporal midpoints of vowels. Lindblom [5] represented Swedish vowels with the values of the first three formants at the time at which the first derivative of the corresponding trajectory was zero. In the study by Di Benedetto [6], the sampling points of formants were chosen at the time at which F_1 reached its maximum. The motivation for this choice was the concave upward shape of the F_1 trajectory of a vowel between two consonants, which is consistent with the prediction of the acoustic theory. In fact, under the shape assumption, the maximum F_1 point is equivalent to the point at which the first derivative is zero. However, the numeric calculation of the derivative amplifies the noise of trajectory measurements, which makes it difficult to determine the zero point automatically from data. This difficulty is more severe for dysarthric speech due to more irregularities. Therefore, in the present study, we chose the maximum point on the F_1 trajectory of each vowel to approximate its stable point.

As for the F_2 trajectory of a vowel, we assumed that it could only be in one of the following four shapes: concave upward or downward, or monotonically increasing or decreasing. When it was in the concave upward or downward shape, we chose the maximum or minimum point as the stable point of the F_2 trajectory. We have observed that the maximum or minimum does not necessarily occur at the same instant of time as the maximum of the F_1 trajectory. This observation could be a consequence of the fact that different articulators can move relatively independently during speech production. When the F_2 trajectory was monotonically increasing or decreasing, we chose the stable point corresponding to the median F_2 value of the trajectory.

In summary, the procedure to measure F_1 and F_2 values at stable points is described as follows. The formant trajectories are extracted by the formant tracker at 10-millisecond intervals. Then a 3-order median filter is used to suppress impulsive noise in the trajectory data. A forward-reverse low-pass filter, whose impulse response is a 5-tap Hanning window, is used to smooth the F_1 and F_2 trajectories. Within the vowel section, the maximum of the F_1 trajectory is then obtained. In order to automatically determine the shape of the F_2 trajectory of the vowel, F_2 data are tested by four types of shape-constrained regressions, including an increasing isotonic regression, a decreasing isotonic regression, a unimodal regression, and a reverse unimodal regression [7]. The shape is determined by the least regression error among the tests. According to the shape of the F_2 trajectory, the stable point and
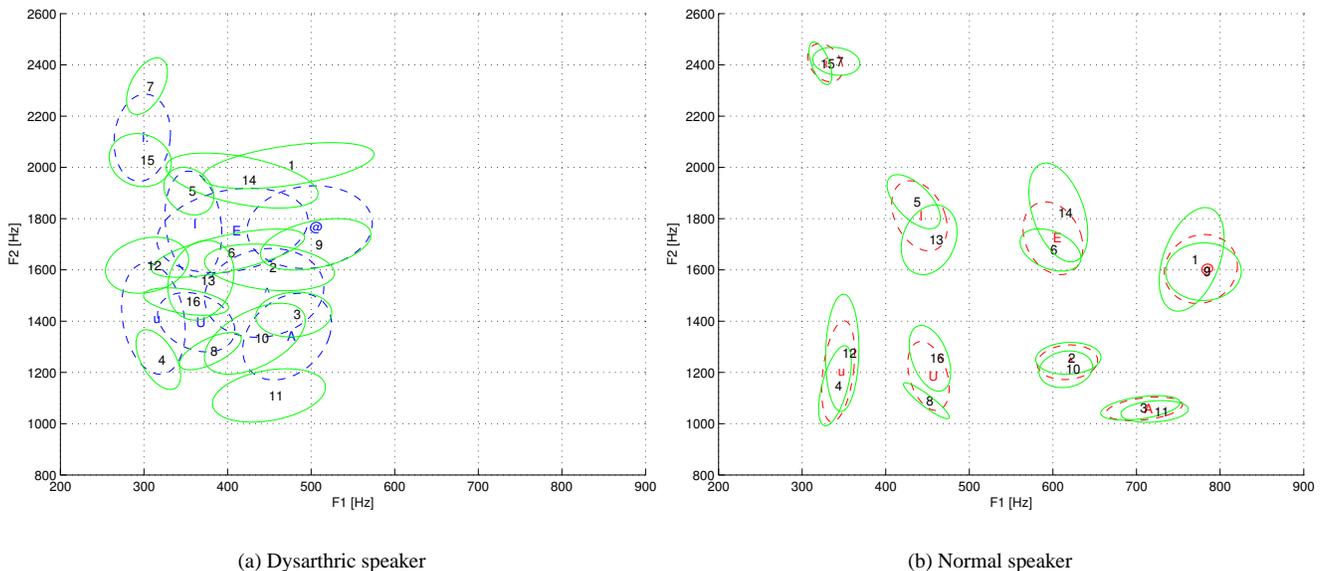
(a) Dysarthric speaker

(b) Normal speaker

Fig. 2. F_1 and F_2 formant frequency targets for the 8 vowel classes of the training set, modeled as Gaussians (blue and red, dashed), and projections of corresponding joint density Gaussian mixture model components (green, solid). Each vowel class is modeled by two components for a total of 16 components. Ellipses are centered at component means and radii are set to equal standard deviations.

F_2 value are determined. The F_1 and F_2 values measured at the stable points are used as the estimates of formant targets during the training and modification process.

2.4. Training

In order to transform formant targets $\mathbf{x}$ of the dysarthric speaker to formant targets $\mathbf{y}$ of the normal speaker, we must train a transformation function $\mathcal{F}$. As in previous work [8], we choose a piecewise linear, probabilistic function, derived from parameters of a Gaussian mixture model (GMM) that models the joint density of $\mathbf{x}$ and $\mathbf{y}$. The predicted normal formant targets are given by

$$\hat{\mathbf{y}} = \mathcal{F}\left(\mathbf{x}\right) = \sum_{q=1}^{Q}\left(\mathbf{W}_q\mathbf{x} + \mathbf{b}_q\right) \cdot p\left(c_q | \mathbf{x}\right) \tag{1}$$

where $\mathbf{W_q}$ and $\mathbf{b_q}$ represent the qth linear transformation, and the term $p\left(c_q|\mathbf{x}\right)$ denotes the GMM posterior probability that the input vector "belongs" to class q.

There are several ways in which the parameters of the GMM can be estimated. For example, the GMM can be trained in *unsupervised* mode using a standard EM algorithm, or it can be trained in *supervised* mode by using the vowel class information directly (in this case $Q = 8$). Initial informal perceptual tests showed that a *semi-supervised* method worked best. In this scheme, we performed a K-means clustering separately on the joint density of each vowel class. After estimating two codewords per class, a two-component sub-GMM is constructed by letting the Gaussian means equal the codewords, and adjusting the priors and the covariances appropriately. We preferred a K-means method over an EM method because the former produces density models that have less "overlap" as compared to the latter, which produces densities that may model the data more accurately but less distinctly (this is also known as the *information-modeling* trade-off [9]).

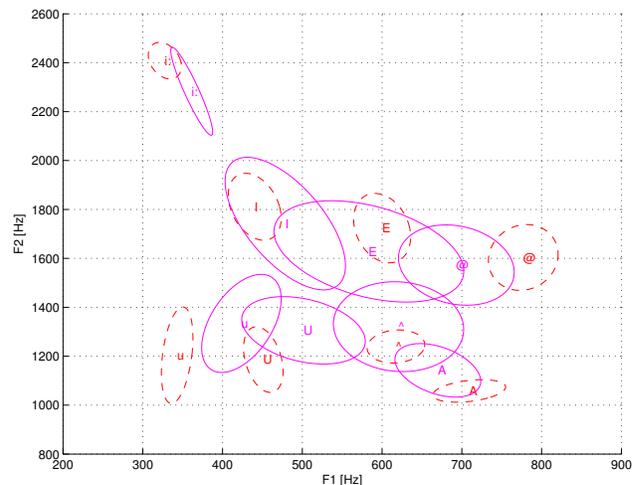Finally, the eight sub-GMMs are assembled into a single 16-

Fig. 3. The transformed vowel space. Shown are Gaussian approximations of formant targets of the normal speaker for the 8 vowel classes (red, dashed) and Gaussian approximations of transformed dysarthric data grouped by their known classes (magenta, solid).

component GMM, as shown in Figure 2. The advantage of using the semi-supervised approach lies in the fact that component allocation is partially guided, but components can still model any non-Gaussian behavior of the data within each class. Figure 3 shows the transformed vowel space.

2.5. Modification
2.5.1. Energy Modification
One of the problems of using energy values of the dysarthric speaker directly is that they result in modified speech with significant

flutter (variations in energy). The variations in energy measurement are probably due to the high levels of "vocal fold fry" [10]. We therefore smooth the energy trajectory with a combination of zero-phase filtering using a normalized Hanning window and median filtering.

2.5.2. F_0 Generation

Generally, F_0 of the dysarthric speaker was characterized by excessive amounts of jitter (variations in F_0), and was perceived as rough. We decided to discard the original F_0 and generate synthetic F_0 values. We have tried three different variations in implementation: 1) let F_0 be directly proportional to the energy; 2) find one or more peaks in the energy trajectory, and consider them as peak locations for accent curves; 3) for CVC contexts, consider 33% of the vowel duration as the center of the peak location of a single accent curve [11].

The first variation is the most simple and produced somewhat natural, albeit meaningless, pitch changes. The second variation attempts to identify regions of emphasis based on energy and uses the superpositional F_0 model to create accent and phrase curves. For the purposes of this paper, we used the third variation, which assumes that the emphasis occurs during the production of the vowel.

2.5.3. Formant Modification

Because the output of the formant tracker was more erratic on dysarthric speech than on the normal speech, we used zero-phase filtering to smooth the formant trajectories prior to any other processing.

The first step consisted of obtaining modified formant target values for the center of the vowel portion (which is a simplification, since these targets may occur in locations other than the center). There were three types of modifications performed, each corresponding to a different condition during the perceptual test described in Section 3:

clean The modified formant frequency targets were set to the *dysarthric* speaker's formant frequency targets, as obtained from the process described in Section 2.3.2

map We *predicted* the normal speaker's formant frequency targets from the dysarthric formant frequency targets using the transformation function introduced in Section 2.4

oracle The modified formant frequency targets were set to the *normal* speaker's formant frequency targets

Since the targets only included F_1 and F_2, we estimated F_3 and F_4 from the average values of the center third of the vowel. The bandwidths $B_1 - B_4$ were set to appropriate constant values. The **clean** and **oracle** conditions can be regarded as the baseline and best possible system performance, respectively; the performance of the **map** condition can be expected to be within this range.

Once the modified target values were available, the modified formant trajectories needed to be calculated. In the absence of a sophisticated method that can properly account for coarticulatory effects [12], we employed a strategy of adding a *synthetic* target trajectory to the originally *observed* trajectories (for both frequencies and bandwidths), weighted by a *crossfade* function. The existence of the crossfade function avoided discontinuities by slowly changing from the observed formant values on the left side of the vowel to the synthetic values, and then back again; this is important since, in contrast to energy and F_0 features which are modified and generated in voiced regions (which includes all sonorants), formants were modified in the vowel region only. We chose the nth root of a Hanning window as crossfade function, where n

can adjust the fade-in and fade-out "speed" (we set $n = 3$). The synthetic trajectory was set to constant values for the entire modification region. Crossfading the observed formant frequencies with constants resulted in trajectories that gave reasonable approximations for concave and convex formant movements.

2.6. Synthesis

The implementation of our synthesizer is very similar to a Klatt formant synthesizer [13]. We use manually adjusted, global values for the source parameters TL (spectral tilt, implemented as a first-order filter) and OQ (open quotient). For each synthetic speech frame, we produced two periods of excitation using the synthetic F_0 values. Based on the first four formant frequency and bandwidth values we constructed the all-pole formant filter. We also constructed a global spectral adjustment filter, implemented as a low-order, linear phase FIR filter, which was tuned so that the normal speaker's spectral balance was reproduced correctly on average. Finally, the excitation waveform was filtered with the combined formant and adjustment filter to render the output frame. The frame's energy was adjusted to the modified energy specifications, and the original high band speech waveform was added to it. To produce the final speech waveform, every frame, unvoiced and voiced, was overlap-added using an asymmetric trapezoidal window to give good transitions between voiced/unvoiced regions as well as between neighboring voiced frames.

3. EVALUATION

Two perceptual experiments were conducted. The first experiment evaluated the intelligibility of the original dysarthric speech, modified dysarthric speech, and original normal speech. The second experiment measured the quality of the original dysarthric speech as compared to the **clean** dysarthric speech.

3.1. Stimulus Types

The five types of stimuli used in these experiments included the **clean**, **map**, and **oracle** modifications to dysarthric speech described in Section 2.5.3, as well as the original waveform of dysarthric speech (referred to as **origDys**) and the original waveform of the normal speaker (referred to as **origNor**).

The **clean** stimuli were used to evaluate the intelligibility and quality of dysarthric speech that has had its energy, F0, and formant structure drastically simplified, but without efforts to modify the speech for improved intelligibility.

The **map** stimuli were used to evaluate the intelligibility of dysarthric speech in which the formant values have been modified in an effort to improve intelligibility. This modification of formants relies on the formant-transformation procedure described in Section 2.4 and, as such, could be implemented in a working system.

The **oracle** stimuli were used to evaluate the intelligibility of dysarthric speech in which the formant values have been modified so that they reach values that are thought to be correct for the purposes of intelligibility. This modification of formants represents the best output of a hypothetical formant-transformation procedure that has been perfectly trained. As such, results from **oracle** stimuli can not yet be implemented in a working system, but illustrate best-possible performance using the current system architecture.

3.2. Perceptual Experiments

The perceptual experiments were conducted with 6 males and 4 females, all of whom were native speakers of American English and
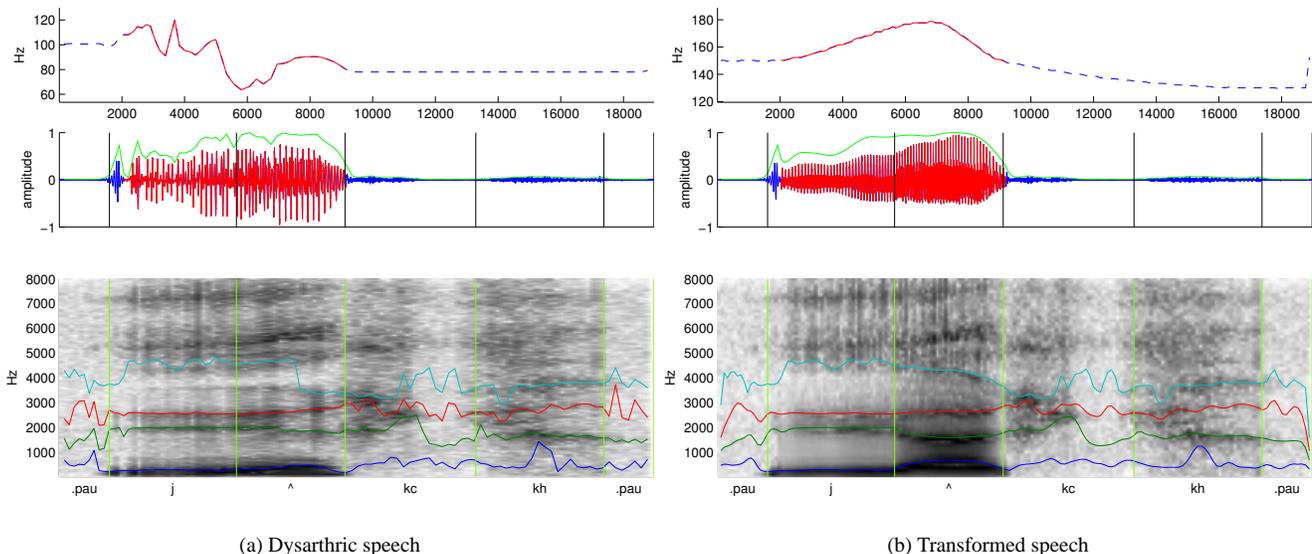
(a) Dysarthric speech   (b) Transformed speech

Fig. 4. Analysis of dysarthric speech and synthesis of transformed speech. The top panel displays F_0 values in voiced (red, solid) and unvoiced (blue, dashed) regions. The middle panel displays the speech waveform (red/blue) and scaled energy values (green). The bottom panel shows the spectrogram superimposed with formant frequencies F_1–F_4.

had no known hearing problems. These subjects were unfamiliar with dysarthric speech. Subjects took these tests using graphical user interfaces that presented the stimuli and possible response choices, and then recorded the responses. Replaying the stimuli was not possible. The stimuli were played over loudspeakers in a quiet room. Two stimuli were presented at the beginning of each test to familiarize the subjects with the procedure; responses from these stimuli were not recorded. The waveforms were normalized for energy levels and the signal was trimmed to have no more than 100 milliseconds of silence at either end.

3.2.1. Intelligibility Testing

The evaluation of intelligibility had a structure similar to that proposed by Kent *et al.* [14] for measuring the intelligibility of speech. The 50 CVC utterances with five stimulus types (**origDys**, **clean**, **map**, **oracle**, and **origNor**) yielded 250 stimuli in total. Each subject evaluated 100 of these stimuli. Each subject listened to the same (random) ordering of the CVCs but a different stimulus type within that ordering, so that the order of presentation had no effect on the relative intelligibility results.

In conducting this test, each CVC was aurally presented. Subjects chose the vowel that they heard from the list of all 8 possible vowels. Vowels were represented using both phonetic symbols (for subjects who were familiar with phonetics) and complete words that contain those vowels (for subjects who were unfamiliar with phonetics). The total time to complete this test was about 15 minutes.

3.2.2. Quality Testing

The goal of the test of speech quality was to evaluate whether or not the analysis of dysarthric speech and subsequent synthesis using smooth F_0, formant, and energy contours would improve the perceived quality of the dysarthric speech. It was hypothesized that such processing would improve perceived quality because of the preponderance of vocal fry in the dysarthric speaker's voice.

The evaluation of speech quality used a standard Comparison Category Rating (CCR) test. In conducting this test, CVC pairs were aurally presented in sequence, using the **origDys** and **clean** stimulus types. The ordering of the presentation of these stimulus types was randomized. Subjects were asked to indicate the change in speech quality of the two speech samples using a response scale, resulting in a comparison mean opinion score (CMOS). The scale included the following rankings (with assigned scores in parentheses): "much worse" (-2), "slightly worse" (-1), "about the same" (0), "slightly better" (+1), and "much better" (+2).

In order to evaluate the change in quality as a function of the presence of vocal fry, the evaluation consisted of 40 utterances containing vocal fry and 10 utterances without vocal fry. The presence of vocal fry was determined using a glottalization detector [15] in the vowel regions of the dysarthric speech. The total time to complete this test was about 15 minutes.

3.3. Results

For the intelligibility test, the average correct response rate for the five stimulus types were: **origDys** = 43.4%, **clean** = 38.5%, **map** = 45.5%, **oracle** = 63.5%, and **origNor** = 99.0%. Results of a planned-comparison one-tailed t-test [16] are displayed in Table 1. The improvements from **clean** to **map** and from **map** to **oracle** are significant at the 5% level. An analysis of the results per target vowel indicates that vowels that should have a high value of F1 had improved intelligibility and vowels that should have a low value of F1 had reduced intelligibility. This is shown in Table 2.

For the test of speech quality, we computed the mean opinion score of the **clean** utterances relative to the corresponding **origDys** utterances. For all utterances, the mean opinion score of the **clean** utterances was 0.19. For those **origDys** utterances considered to have vocal fry, the mean opinion score of the **clean** utterances was 0.26. For utterances without vocal fry, the mean opinion score was -0.08. Results of t-tests between the **origDys** and **clean** utterances are shown in Table 3.

Comparison	Diff. (%)	t-test	p value
origDys *vs.* **clean**	-4.9	-0.760	N.S.
origDys *vs.* **map**	2.1	0.293	N.S.
clean *vs.* **map**	7.0	1.982	0.039
map *vs.* **oracle**	18.0	3.756	0.002

Table 1. Results of planned-comparison one-tailed significance testing of intelligibility.

Vowel (Word)	OrigDys (%)	clean (%)	map (%)	oracle (%)	origNor (%)
i: (beet)	82	71	75	83	100
I (bit)	73	71	54	82	100
E (bet)	28	33	25	38	100
@ (bat)	29	4	54	71	100
u (boot)	56	50	4	54	100
U (book)	44	50	42	42	93
^ (but)	31	8	33	33	100
A (Bob)	13	18	71	96	100

Table 2. Results of intelligibility test on a per-vowel basis

4. CONCLUSION

In conclusion, these preliminary results show that a method that modifies formants without knowledge of vowel identities may be capable of enhancing intelligibility. The major issue that needs to be addressed is why the **clean** condition results in a decrement in intelligibility and an increment in quality. Results indicate that the increment in quality is due to the removal of vocal fry. This follows not only from the overall result, but also from the difference in results for stimuli with versus without vocal fry. The decrement in intelligibility may have multiple causes, and include the following. First, the consonants were re-synthesized fairly crudely; this will be addressed by using a mixed formant/sinusoidal re-synthesis process. Second, the formant trajectories were drawn using a scheme that leaves the formant values at the vowel boundaries intact. One problem that this generates is unusually strong formant movement. New schemes will be considered that modify the formant trajectories well into the consonants. Additionally, further control experiments need to be conducted in which we compare the cleaned version with a similar version in which the original formants are used. Although the mapping operation was clearly successful, performance on the /E/, /u/, and /I/ vowels was problematic. It is important to point out that the mapping operation does not reduce overlap in the productions of different vowels. Instead, it moves the vowel space such that more vowels are in the appropriate regions of the vowel space. Because of the relatively high degree of overlap, formant mapping unavoidably creates errors. Two avenues of research need to be explored here. One is to obtain a more accurate *perceptual* picture of the formant space, and use this picture instead of the normal speaker's productions as targets. The other is to reduce overlap by attempting to "de-coarticulate" the vowels. Work is underway on precisely this issue [12]. Finally, we want to remind ourselves of the fact that this speaker represents one of many patterns of dysarthria, and that we should be prepared for deeply different methods being required for different syndromes.

Utterance Type	Diff. Score (-2 to +2)	t-test	p value
All Utterances	0.190	2.571	0.015
Vocal Fry Only	0.259	2.759	0.011
Non-Fry Only	-0.079	-1.299	N.S.
Fry *vs.* Non-Fry	0.338	2.858	0.009

Table 3. Results of significance testing for speech quality change

5. REFERENCES

[1] Electronic Speech Enhancement Inc., "The Speech Enhancer," http://www.speechenhancer.com.

[2] J. L. Hieronymus, "ASCII phonetic symbols for the worlds languages: Worldbet," *Bell Labs Technical Memorandum*, 1993.

[3] Entropic Research Laboratory, Inc., *Entropic Signal Processing System (ESPS) / waves+*, August 1993, Version 5.0.

[4] K. N. Stevens and A. S. House, "Perturbation of vowel articulations by consonantal context: An acoustical study," *Journal of speech and hearing research*, vol. 6, pp. 111–128, 1963.

[5] B. Lindblom, "Spectrographic study of vowel reduction," *Journal of the acoustical society of america*, vol. 35, no. 11, pp. 1773–1781, Nov. 1963.

[6] M.-G. Di Benedetto, "Vowel representation: Some observations on temporal and spectral properties of the first formant frequency," *Journal of the acoustical society of america*, vol. 86, no. 1, pp. 55–66, 1989.

[7] R.E. Barlow, D.J. Barholomew, J.M. Bremner, and H.D. Brunk, *Statistical Inference under Order Restrictions*, Wiley, Chichester, 1980.

[8] Alexander Kain and Mike Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings of ICASSP '98*, Seattle, WA, May 1998, vol. 1, pp. 285–288.

[9] Michael Kearns, Yishay Mansour, and Andrew Y Ng, "An information-theoretic analysis of hard and soft assignment methods for clustering," in *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*. 1997, pp. 282–293, Morgan Kaufmann.

[10] Ingo R. Titze, *Principles of Voice Production*, Prentice-Hall, Inc., 1994.

[11] J. van Santen and B. Möbius, "A quantitative model of F_0 generation and alignment," in *Intonation — Analysis, Modelling and Technology*, Antonis Botinis, Ed., pp. 269–288. Kluwer, Dordrecht, 2000.

[12] X. Niu and J. P. H. van Santen, "A formant-trajectory model and its usage in comparing coarticulatory effects in dysarthric and normal speech," in *Models and analysis of vocal emissions for biomedical applications (MAVEBA), 3rd international workshop*, 2003, pp. 233–236.

[13] Dennis H Klatt, "Review of text-to-speech conversion for english," *JASA*, vol. 82, no. 3, pp. 737–793, 1987.

[14] Kent, R.D., Weismer, G. Kent, J.F., and Rosenbek, J.C., "Toward phonetic intelligibility testing in dysarthria," *Journal of Speech and Hearing Disorders*, vol. 54, pp. 482–499, Nov. 1989.

[15] John-Paul Hosom, *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information*, Ph.D. thesis, Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, Beaverton, OR, USA, May 2000, Published as Technical Report CSE-00-TH-002.

[16] Winer, B.J. and Brown, D. R and Michels, K.M., *Statistical Principles in Experimental Design*, McGraw-Hill, Inc., New York NY, 1991.