

Improving the intelligibility of dysarthric speech

Alexander B. Kain^{a,*}, John-Paul Hosom^a, Xiaochuan Niu^a, Jan P.H. van Santen^a,
Melanie Fried-Oken^b, Janice Staehely^b

^a Center for Spoken Language Understanding, OGI School of Science & Engineering, Oregon Health & Science University, Portland, OR, USA

^b Departments of Neurology and Otolaryngology, Oregon Institute on Disability and Development, Oregon Health & Science University, Portland, OR, USA

Received 29 July 2006; received in revised form 18 April 2007

Abstract

Dysarthria is a speech motor disorder usually resulting in a substantive decrease in speech intelligibility by the general population. In this study, we have significantly improved the intelligibility of dysarthric vowels of one speaker from 48% to 54%, as evaluated by a vowel identification task using 64 CVC stimuli judged by 24 listeners. Improvement was obtained by transforming the vowels of a speaker with dysarthria to more closely match the vowel space of a non-dysarthric (target) speaker. The optimal mapping feature set, from a list of 21 candidate feature sets, proved to be one utilizing vowel duration and F1–F3 stable points, which were calculated using shape-constrained isotonic regression. The choice of speaker-specific or speaker-independent vowel formant targets appeared to be insignificant. Comparisons with “oracle” conditions were performed in order to evaluate the analysis/re-synthesis system independently of the transformation function.

© 2007 Elsevier B.V. All rights reserved.

Keywords: Speech processing; Speech transformation; Speech modification; Intelligibility; Dysarthria

1. Introduction

Dysarthria refers to a group of speech disorders resulting from disturbances in muscular control over the speech mechanism due to damage of the central or peripheral nervous system. Individuals with dysarthria have problems in oral communication due to paralysis, weakness, or incoordination of the speech musculature (Darley et al., 1969). This neurogenic motor speech impairment is associated with diseases and conditions that are chronic or long-term (Yorkston et al., 1999). Some neurological conditions, such as traumatic brain injury or stroke, produce a non-progressive dysarthria, while others, such as Huntington’s disease, Parkinson’s disease or amyotrophic lateral sclerosis (ALS), produce a degenerative dysarthria that degrades speech over time. Dysarthria can be described as an impairment

in one or more of the processes of speech production: respiration, phonation, resonance, articulation, and prosody. The disorder of movement is due to abnormal neuromuscular execution that may affect the speed, strength, range, timing, or accuracy of speech movements (Duffy, 2005). For example, an individual with ataxic dysarthria secondary to Friedrich’s ataxia may present with a slow rate for individual and repetitive speech movements, an excessive range of speech movements, reduced muscle tone and irregular speech rhythm. The patient’s speech may include imprecise consonants and distorted vowels, irregular articulatory breakdowns, excessive or equal stress to all syllables, and a slow rate of speech with a phonatory-prosodic insufficiency described as harsh, monotonous and monoloudness. Individuals with Parkinson’s disease usually present with a hypokinetic dysarthria. They complain that their voice is quieter or weak, their speech rate is too fast, words are indistinct and that it is often difficult to get speech started. Persons with multiple sclerosis, on the other hand, may complain of slurred speech with changes in

* Corresponding author. Tel.: +1 503 329 9604; fax: +1 503 748 1306.
E-mail address: kain@cslu.ogi.edu (A.B. Kain).
URL: www.cslu.ogi.edu (A.B. Kain).

pitch, hypernasal speech, hoarseness and poor loudness control.

Very little data are available on the estimated incidence or prevalence of people with dysarthria in the United States. Yorkston et al. (1988) contend that motor speech disorders probably represent a significant proportion of the communication disorders seen in medical speech-language pathology practices. Duffy (2005) found that 36.5% of speech pathology diagnostic consultations from the Department of Neurology at the Mayo Clinic from 1987 to 1990 were motor speech disorders. Another study of 77 patients with multiple sclerosis showed that 51% had mild to severe dysarthria (Hartelius et al., 2000). These percentages may be extrapolated to the larger population of individuals with neurologic disease who have required speech pathology evaluations in the past.

The type of treatment for persons with dysarthria is dependent on a number of circumstances, including diagnosis, severity of dysarthria, present needs and prognosis, environment, and the speaker's acceptance or motivation for speaking. Generally, treatment options range from behavioral techniques to speech generating devices that replace the patient's speech.

Speech generating devices fall within the clinical field of augmentative and alternative communication (Beukelman and Mirenda, 2005). Certainly dysarthric speakers prefer to use their own natural skills to express thoughts. However, listeners prefer synthetic speech over natural dysarthric speech, if the latter has low intelligibility (Drager et al., 2004). Often a dysarthric speaker will use natural speech and a device together, depending on the message, listener, environment, and importance of the interaction. For example, a person with moderate dysarthria secondary to a stroke might use a speech generating device to converse over the telephone with his children, but rely on dysarthric speech in a romantic dinner setting.

Devices are developed with three critical features in mind: the type of user input (e.g. keyboard, pointers, or head movements), the type of output (e.g. play-back of recorded speech or synthesized speech), and the type of language representation, which is the form of symbols that will be used to represent thought and language (e.g. photographs/pictures (Words+ Inc.; Visser), alphabet, or symbols (Prentke Romich Company; Semantic Compaction Systems)). While speech generating devices can replace or supplement natural speech, they have not yet been designed to mimic the speed or ease of production that is experienced by oral communicators.

To date, only one device exists on the commercial market that attempts to improve the natural dysarthric speech itself. The *Speech Enhancer* device (Electronic Speech Enhancement Inc) claims to "clarify" dysarthric speech; however, no published controlled research is available. The exact nature of their electronic algorithm is proprietary. We speculate that it is an adaptive equalizer that continuously adjusts the gain in several frequency bands to fit a target spectral envelope, in order

to amplify those regions of speech most relevant to speech perception.

In this article, we report on our recent research with the goal of enabling people with dysarthria to be understood by the general population. Our approach is to *transform* the original dysarthric speech signal by performing a detailed speech *analysis*, applying a trained *transformation function* to the obtained speech features, and to synthesize a new speech signal from the transformed speech features. The resulting speech signals have been used as stimuli in a formal perceptual listening test to measure the efficacy of our method. In the present work, we aim to answer three questions: (1) Can vowels from a dysarthric speaker be transformed for improved intelligibility, (2) How do speaker-specific or speaker-independent vowel formant targets affect the intelligibility of transformed speech, and (3) What is an appropriate feature set for both input and output of the transformation function.

The rest of the article is organized as follows: Section 2 reviews our preliminary work on this subject. Section 3 introduces the key ideas of our approach and discusses the speech corpus (Section 3.2) as well as the analysis (Section 3.3), transformation (Section 3.4), and synthesis (Section 3.5) of speech in detail. The design, administration, and results of an evaluation of our method are discussed in Section 4, and we conclude with Section 5.

2. Preliminary experiments

In a first experiment (Hosom et al., 2003), we studied the effects of modifications to the dysarthric speech signal. Specifically, we measured the contributions of prosody and short-term spectra of certain speech sounds to the intelligibility of sentences. Using "hybrid" stimuli created from parallel natural and dysarthric speech recordings of nonsense sentences (Menéndez-Pidal et al., 1996), we found that replacing the short-term spectrum of a dysarthric speaker (identified as LL) with a non-dysarthric speaker's short-term spectrum (while keeping the prosody of the dysarthric speaker) led to an intelligibility of 87% as compared to a baseline of 68%. Moreover, replacing just vowels, liquids, and glides (VLG) led to an intelligibility of 75%, whereas replacing all non-VLG speech sounds resulted in an intelligibility of 73%. Replacing the prosody (pitch, duration, and energy) of the utterance with that of a non-dysarthric speaker, while keeping the dysarthric short-term spectrum, led to an intelligibility of 75%. These results demonstrated the potential of improving the intelligibility of dysarthric speech by a partial modification of the speech signal, even though the factors that influence intelligibility are both numerous and co-dependent.

A similar approach was taken by Maassen and Povel, who modified deaf speech with the intent to increase its intelligibility (Maassen and Povel, 1984). Replacing the fundamental frequency trajectory with an artificial one yielded a 7% improvement (from 20% to 27% words correctly identified). They also experimented with replacing

phoneme durations along with the fundamental frequencies with values from hearing speakers, resulting in a larger improvement (from 24% to 34%) (Maassen and Povel, 1985). However, a segmental replacement caused the most dramatic increase, from 24% to 72%. The researchers concluded that, for deaf speech, improving articulation is more important than improving prosody.

From a desire to begin the design and implementation of the transformation system with a single, simple, yet effective component, we decided to initially focus on transforming vowel short-term spectra (and later durations). However, it remained to be seen whether a transformation function could be trained that would transform unintelligible dysarthric vowels (in a CVC context, from a second speaker identified as DC) into intelligible ones. To answer this question, we created a system which analyzed and transformed vowel formant frequencies F1 and F2, and then used formant synthesis to render the final output (Kain et al., 2004). Unfortunately, the system did not show a significant improvement in intelligibility under testing conditions. When taking a closer look at specific instances of perceptual test errors (when listeners selected the wrong vowel), we could identify problems involving incorrect duration, errors in formant tracking (specifically F1 near consonants), and the synthetic formant trajectory model. These observations led to the transformation system and evaluations presented here.

The choice to initially focus on vowels is supported by various research in the dysarthria literature that links vowel articulatory deficits to reduced speech intelligibility. For example, an analysis of the phonetic impairments of 25 male individuals with ALS revealed that the most disrupted phonetic features included voicing contrast, nasalization, and regulation of tongue height for vowels (Kent et al., 1990). In another study, it was shown that individuals with dysarthria exhibited smaller vowel space areas, as compared to a control group, and that the vowel space area accounted for 45% of the variance in speech intelligibility (Turner et al., 1995). Examining the types of errors that occurred during intelligibility testing, it was found that vowels had a significantly greater contribution to the average rate of incorrect judgements than consonant items (Ziegler et al., 1988). Finally, evidence from X-ray microbeam data suggests that some dysarthric speakers present with subtle disruptions of coordinated articulatory behavior, as observed during the production of the vowel /u/ (Weismer et al., 2003).

3. Method

3.1. Overview

The key idea of our approach is to improve intelligibility by analysis, transformation, and synthesis of a small set of perceptually-relevant speech features. The transformation step consists of mapping the dysarthric features towards known good target features by means of a trained transfor-

mation function. The particular choice of speech features is motivated by the need to represent speech intelligibly, but not necessarily very naturally or with the dysarthric speaker's own voice. At the same time, the number of training parameters should be kept small to allow training of the transformation function with a relatively small amount of training data. For these reasons, the proposed speech features and synthesis method are similar to a formant speech-synthesis approach, producing highly intelligible and controllable speech from a compact representation. We are encouraged by initial results from research in which the formants of disordered speech were modified manually (Qi et al., 1995, 1996).

Fig. 1 shows a flowchart of the proposed system's architecture. The speech material in the recording database consisted of speech signal waveforms, voicing values, and consonant-vowel-consonant (CVC) boundaries (but not their identities). In this proof-of-concept stage, it was assumed that automatic and sufficiently accurate voicing and phoneme-boundary detection systems can be developed in the future; our work focused on the main problem of improving vowel intelligibility. The system processed speech on a frame-by-frame basis, pitch-synchronously. An entire utterance was analyzed to obtain formant frequency values, to be used as input to the formant

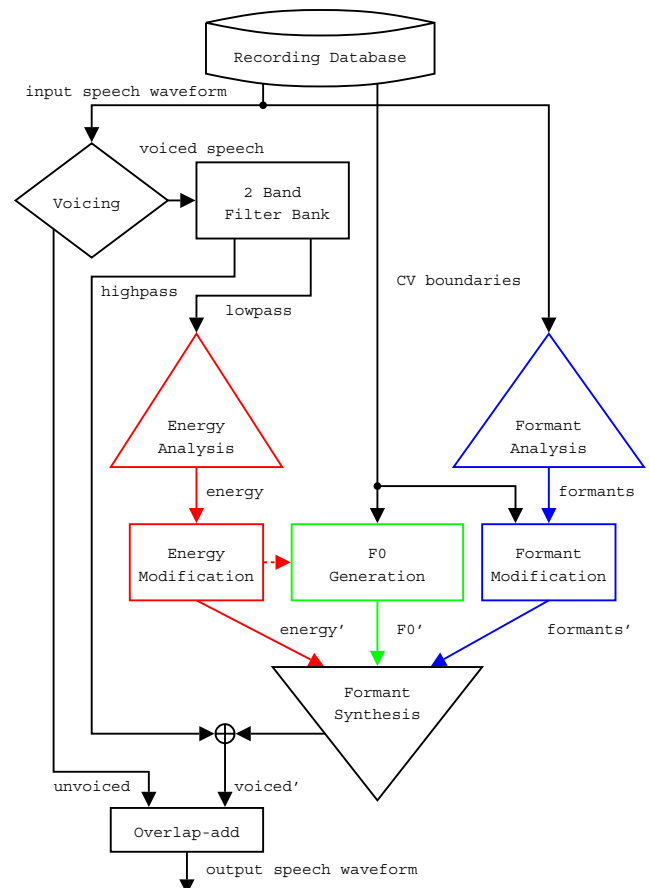


Fig. 1. System architecture for improving intelligibility of dysarthric speech.

modification operation. Unvoiced frames of speech, however, were passed directly to the output. Additionally, any speech frequencies above 4 kHz were passed through, unmodified, to the output signal.

To generate the voiced regions of the transformed speech, the system performed modifications to the energy and formant features, and generated a new F0 trajectory from the CVC boundary information. Energy modification was applied because the dysarthric speech often contained significant energy flutter (variations in energy), likely caused by high levels of “vocal fry” (Titze, 2000). Similarly, the F0 trajectory of a dysarthric speaker often contained significant jitter (variations in F0). In our experiments, we discarded the original F0 values, which were often estimated with large errors, in favor of a synthetic F0 contour, generated by a simple superpositional intonation model (van Santen and Möbius, 2000). Formants were modified by estimating a representative formant vector for the entire dysarthric vowel, mapping that vector to the transformed formant vector using a nonlinear, trained function, and then creating the transformed formant trajectory. Finally, vowel durations were modified.

It should be noted that our method did not employ other, simpler modifications to the signal, such as dynamic volume compression/expansion schemes. Even though these modifications are commonly believed to improve the intelligibility of the speech signal (and are thus commonly used in simple speech enhancement devices for persons with dysarthria and other speech processing areas such as broadcasting), the authors wanted to measure improvements that were only a consequence of the proposed method.

3.2. Database design and recording

In order to deal with the formidable complexity of the task, we limited ourselves to studying CVC contexts from a special-purpose database. We also used an additional database for analysis of vowel targets.

3.2.1. CVC Database

The vowels in the CVC “words” consisted of 4 front vowels (/i:/, /I/, /E/, and /@/, using Worldbet symbols for transcription (Hieronymus) and 4 back vowels (/u/, /U/, /^/, and /A/). These vowels represent typical vocal-tract configurations in American English. The vowel />/ was omitted because in West-Coast American English this vowel is often identical to /A/. Diphthongs were omitted because of their dynamics. 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/). The /j/ and /w/ approximants occurred only in the initial consonant position, as they are never syllable-final phonemes in English. The /l/ approximant occurred only in the initial consonant position because word-final /l/ causes a high degree of “coloring” of the preceding vowel. These consonants cover a variety of places of articulation and manner

of articulation among the consonants of English. Nasal consonants were omitted because they may cause nasalization of the neighboring vowel, which makes formant tracking more difficult. The list of CVC words was constructed by randomly generating a unique CVC combination and adding this CVC word to the final list of words only if that CVC word occurred in a 125,000-word pronunciation dictionary of American English (Carnegie Mellon University, 2004). The final number of vowel occurrences is shown in Table 1.

The speech data were utterances from one dysarthric speaker (speaker DC) and one non-dysarthric speaker (speaker JH). The dysarthric speaker was a female native speaker of American English presenting with Friedrich’s ataxia. Clinically, she was judged to be about 70% intelligible and her moderately dysarthric speech was characterized by vowel distortions, monostress and monoloudness, hypernasality, reduced vocal range with a harsh, strangled quality to connected speech. The non-dysarthric speaker was a male native speaker of American English. Each speaker read 278 isolated monosyllabic CVC words. The speech signal was recorded directly to a hard drive using a head-mounted AKG HSC 200 electret microphone and MAudio Delta 1010 A/D converter. Waveforms were recorded and stored in 16 kHz, 16-bit PCM format. Recordings were made in a quiet, but not acoustically dampened, room. The recording interface for the dysarthric speaker (who was not familiar with phonetic alphabets) consisted of (a) the CVC word, presented on the screen, (b) a word that rhymed with the CVC word, presented on the screen, and (c) the play-back of a previously recorded sample of this CVC from the non-dysarthric speaker. The dysarthric subject then spoke the CVC word after a short tone prompt. Recording of all 278 words by the dysarthric speaker was performed in a single, 1.25 hour session with several rest breaks.

After recording, each utterance from both the non-dysarthric and dysarthric speaker was manually segmented into a sequence of phoneme labels with time alignments.

The CVC database was prepared into several rotations of test and training sets. First, all words from the corpus for both the non-dysarthric and the dysarthric speaker were analyzed (see Section 3.3). Then, the data was split into training (214 feature vectors) and testing (64 feature vectors) data sets. Care was taken to ensure that the testing data set contained a uniform distribution of all available vowels (8 occurrences of all 8 vowels). Actual assignments were performed by random permutation. By seeding the random number generator with 10 different starting states, we obtained 10 different versions (rotations) of splitting the corpus data into training and testing partitions.

Table 1
Number of vowel occurrences in the speech corpus

Vowel	/i:/	/I/	/E/	/@/	/u/	/U/	/^/	/A/
Total	39	42	38	38	31	18	31	41

3.2.2. Vowel-target database

In addition to the 278 CVC words from the dysarthric speaker, nine words were recorded several times at recording sessions over a period of several months. These words were selected in order to elicit formant values that reached their intended targets with minimal coarticulatory influence. These nine words were (with pronunciation indicated in parentheses): “he” (/h i:/), “hit” (/h I t/), “heck” (/h E k/), “hack” (/h @ k/), “who” (/h u/), “hook” (/h U k/), “huff” (/h ^ f/), “ha” (/h A/), and “hoe” (/h oU/).

3.3. Analysis

Energy, voicing, and formant (F1–F3) features were derived using the ESPS *Waves+* software package. We specifically tuned the *getformant* parameters (*preemphasis* = 0.99, *lpcorder* = 20; all other parameters were at their default settings) to avoid making F1 errors at the edges of the vowel. Even so, errors in both formant frequency and bandwidth estimation occurred at times. We decided to not correct these values, in order to demonstrate that the proposed algorithm works with imperfect, but *automatic* formant estimation. Additionally, we measured the initial and final formant slopes by taking the average of the first order difference in the initial third and final third of the vowel region.

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 stable point or section in previous studies of the formant characteristics of vowels. Stevens and House (1963) studied formant values at temporal midpoints of vowels. Lindblom (1963) 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 a study by Di Benedetto (1989), the sampling points of formants were chosen at the time at which F1 reached its maximum. The motivation for this choice was the concave upward shape of the F1 trajectory of a vowel between two consonants, which is consistent with the prediction of acoustic theory. In fact, under the shape assumption, the maximum F1 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 reliably 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 F1 trajectory of each vowel to approximate its stable point.

As for the F2 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 down-

ward shape (as determined by shape-constrained regression, explained below), we chose the maximum or minimum point as the stable point of the F2 trajectory. We have observed that the maximum or minimum does not necessarily occur at the same instant of time as the maximum of the F1 trajectory. This observation could be a consequence of the fact that different articulators can move relatively independently during speech production. When the F2 trajectory was monotonically increasing or decreasing, we chose the stable point as the median F2 value of the trajectory.

The procedure to measure F1 and F2 values at stable points was as follows. The formant trajectories were extracted by the formant tracker at 10 ms intervals. Then a third-order median filter was used to suppress impulsive noise in the trajectory data. A forward–reverse low-pass filter, whose impulse response was a five-tap normalized Hanning window, was used to smooth the F1 and F2 trajectories. Within the vowel section, the maximum of the F1 trajectory was then obtained. In order to automatically determine the shape of the F2 trajectory of the vowel, F2 data were 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. The shape was determined by the least regression error among the tests. According to the shape of the F2 trajectory, the stable point and F2 value were determined. The F1 and F2 values measured at the stable points were used as the estimates of formant targets during the training and modification process.

The four shape-constrained regressions worked as follows. Given a sequence of numbers $\{Y_i\}_{i=1}^N$, the increasing isotonic regression found $\{\hat{Y}_i\}_{i=1}^N$ to minimize $\sum_{i=1}^N (Y_i - \hat{Y}_i)^2 / N$, subject to the monotonicity restriction, $\hat{Y}_1 \leq \hat{Y}_2 \leq \dots \leq \hat{Y}_N$. The problem was solved with the pool-adjacent-violators algorithm (Barlow et al., 1980):

- (1) Start with Y_1 and move to the right until $Y_i > Y_{i+1}$. Pool Y_i and the adjacent Y_{i+1} , by replacing them both with their average Y_i^* .
- (2) Check whether $Y_{i-1} \leq Y_i^*$. If not, pool $\{Y_{i-1}, Y_i, Y_{i+1}\}$ into their average. Continue to the left until the monotonicity requirement is satisfied. Then proceed to the right. The final solutions are $\{\hat{Y}_i\}_{i=1}^N$.

The decreasing isotonic regression performed a similar operation to the above algorithm with the opposite monotonicity restriction and violator check. It found a monotonically non-increasing solution.

The unimodal regression assumed each point of the input sequence could be the unique peak. It applied the increasing regression to the left part of the sequence and the decreasing regression to the right part of the sequence, and calculated the total mean-square error of the estimation. The peak with the least mean-square error was chosen to be the solution. The reverse unimodal regression searched for a unique valley in a similar way by using a

decreasing, and then an increasing regression of the input sequence.

The four regressions were used to estimate four smoothed F2 trajectories. The root-mean-square (RMS) error between each of the estimated F2 trajectories and the original one was calculated as an indicator of the goodness of fit. The shape with the least RMS error was chosen. As a by-product, the RMS errors of the four shapes were used as a 4-dimensional feature (called *F2rms* in some of the configurations in Table 2) that provided one type of description of the shape of the original F2 trajectory. Fig. 2 shows scatter plots of the stable-point F1 and F2 values for all CVC utterances in the corpus (with the vowel identified by the corresponding vowel symbol), for both the dysarthric and non-dysarthric speaker. The relative collapse of the vowel space observed for the dysarthric speaker with Friedrich's Ataxia, as compared to the non-dysarthric speaker, is consistent with the trend identified in the dysarthria literature. F3 stable point values were calculated in the same manner as F2 stable point values.

3.4. Transformation

3.4.1. Input and output features

As stated earlier, we have chosen to transform speech features that are similar to those found in formant synthesis. In a previous study, we used stable points of F1 and F2 for both input (dysarthric speech features) and output (transformed speech features) of the transformation function (Kain et al., 2004). However, further experimentation

had shown that the addition of F3 and vowel duration improved vowel intelligibility (though not technically required in the absence of /9r/). Therefore, we used formant frequencies F1, F2, F3, and vowel duration as the output feature set in the present study. During synthesis (described in Section 3.5), these output features, together with energy and pitch trajectories, completely specified how to synthesize the vowel portion.

As for the input feature set, features in addition to F1 and F2 may improve classification accuracy, but it is unclear which additional features should be used. To discover an appropriate input feature set, given the specific type and amount of available data, we considered a number of different configurations, shown in Table 2. Instead of studying all possible permutations of types of features, we selected configurations that systematically test the performance of individual features. In this table, configurations 1–2 and 5–6 compared simply taking the median of the formant trajectories (*F1median*, *F2median*, and *F3median*) with their respective stable points (*F1stable*, *F2stable*, and *F3stable*), where set 5–6 included F3 information. Configurations 3–4 and 7–8 added duration information. Configurations 9–20 were experiments designed to answer the question of whether the addition of contextual information to the feature set improves performance. Specifically, configurations 9–12 added formant-slope information, where *F1slopeLeft* was defined as the median of the first difference of the F1 formant trajectory during the first and last 30% of the vowel region, and other slope parameters were defined similarly. Configurations 13–16 expressed context

Table 2
Sets of features used as input to the transformation function

Set	Features
1	F1median + F2median
2	F1stable + F2stable
3	F1median + F2median + duration
4	F1stable + F2stable + duration
5	F1median + F2median + F3median
6	F1stable + F2stable + F3stable
7	F1median + F2median + F3median + duration
8	F1stable + F2stable + F3stable + duration
9	F1stable + F2stable + F3stable + duration + F1slopeLeft + F1slopeRight
10	F1stable + F2stable + F3stable + duration + F2slopeLeft + F2slopeRight
11	F1stable + F2stable + F3stable + duration + F1slopeLeft + F1slopeRight + F2slopeLeft + F2slopeRight
12	F1stable + F2stable + F3stable + duration + F2slopeRight
13	F1stable + F2stable + F2rms
14	F1stable + F2stable + duration + F2rms
15	F1stable + F2stable + F3stable + F2rms
16	F1stable + F2stable + F3stable + duration + F2rms
17	F1stable + F2stable + F2poly
18	F1stable + F2stable + duration + F2poly
19	F1stable + F2stable + F3stable + F2poly
20	F1stable + F2stable + F3stable + duration + F2poly
21	F1stable + F2stable + F3stable + duration + energy

F1, F2, F3 refer to the first, second, and third formant frequency, estimated either by the *median* of the trajectory of the center 60% of the vowel region, or by the *stable* point analysis described in the text. *Duration* represents the vowel region duration. *slopeLeft* and *slopeRight* refer to the median first difference in formant frequency value during the first and last 30% of the vowel region. *F2rms* refers to a 4-tuple of values calculated by the procedure described in the text. *F2poly* is a 2-tuple of the first and second coefficients of a second degree polynomial fitted to the F2 trajectory over the vowel region. Finally, *energy* refers to the median energy over the vowel region.

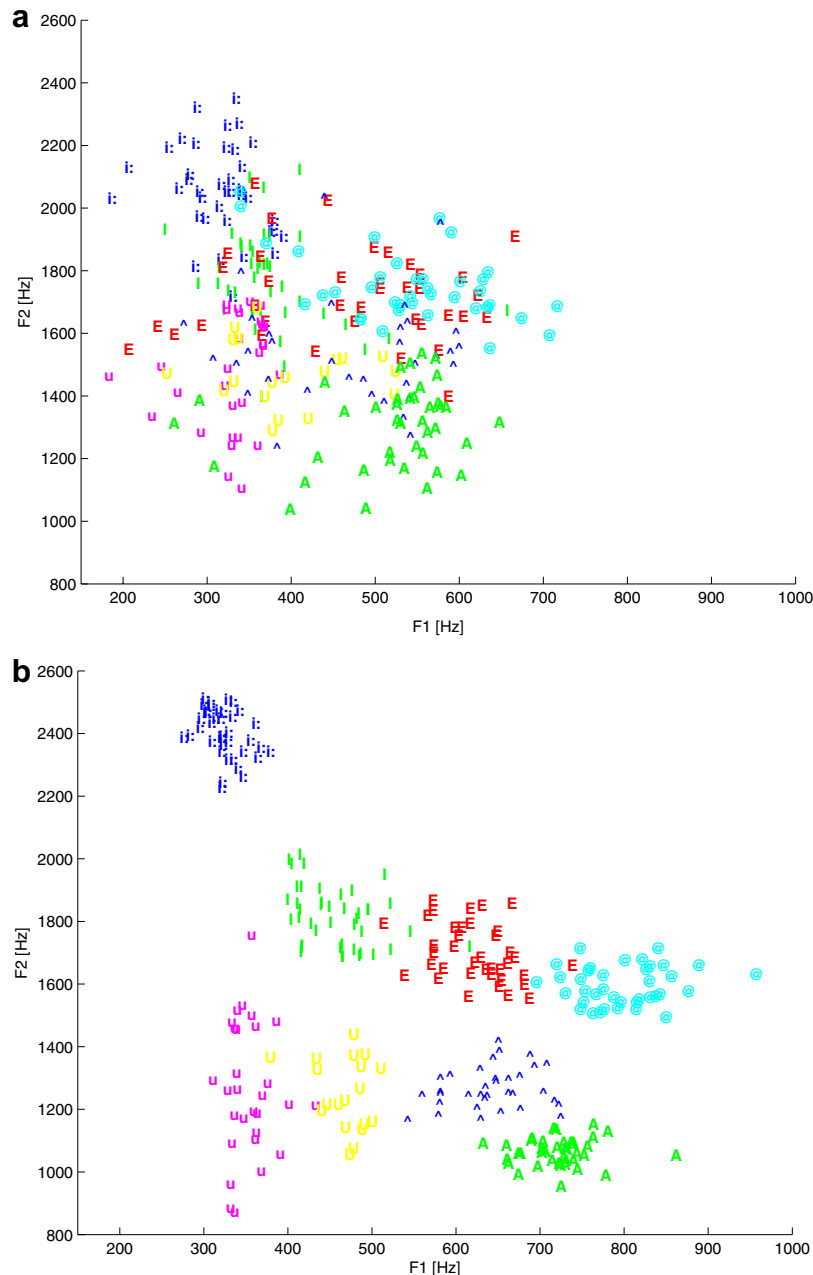


Fig. 2. Scatter plots of the stable-point F1 and F2 values for all CVC utterances in the corpus, with the vowel identified by the corresponding vowel symbol, for both the dysarthric (top panel) and non-dysarthric speaker (color online). (a) Dysarthric speech and (b) non-dysarthric speech.

by adding the F2 distances (called $F2rms$, a 4-dimensional vector) of the vowel to the four template curves, as described at the end of Section 3.3. Configurations 17–20 included context by fitting a second-order polynomial to the formant trajectory and adding the polynomial coefficients as a 2-dimensional vector $F2poly$. Finally, configuration 21 added vowel energy to the best performing configuration from configurations 1–20, which was configuration 8. The best performing set was selected by an objective-evaluation criterion described in Section 3.4.3.

Having considered the types of features to be used as input and output to the transformation function, we now discuss the origin of the data itself. As input, we used the

training dataset of the dysarthric speaker, as described in Section 3.2. As for output targets, we used vowel-specific, context-independent target values; of these, the formant frequency values were either *generic* or *individual*. The generic values were derived from the average female values of Peterson and Barney's seminal work (Peterson and Barney, 1952). Individual values were more difficult to determine, because the subject's dysarthria often prevented intelligible speech. Individual values were derived from numerous recordings made by the subject of the nine vowel-evoking words in the vowel-target database (see Section 3.2.2). Several recordings of each word were listened to, in order to find the most acceptable rendition. Formant targets were

estimated from manual inspection of the spectrogram at the center of the vowel of the selected recording. Then, a formant-synthesis program was used to synthesize vowels from these estimated values. The formant frequency values were then iteratively modified, in small increments, according to the following three criteria: (1) no formant frequency was changed more than 200 Hz from its original estimated value, (2) formants were changed until each synthetic vowel was considered a completely intelligible rendition of its target vowel by two of the authors, and (3) as few formant changes as possible were made when meeting the first two criteria. Table 3 lists the specific formant frequency target values. We note that estimating intelligible output targets from dysarthric speech may not always be possible in the general case.

For target values of vowel duration values, we considered using the average non-dysarthric speaker vowel durations directly, or these durations adjusted for the difference in speaking rate between the two speakers. The speaking-rate difference could be measured from the average ratio of (a) just the vowel durations, (b) all phone durations, or (c) whole word durations. Analysis of the corpus revealed that all three of these ratios were close to 1.8, i.e. the dysarthric speaker spoke almost twice as slowly as the non-dysarthric speaker. However, informal perceptual tests showed that preserving relative consonant-to-vowel duration ratios within a speaker did not contribute to improved vowel intelligibility. Therefore, we chose to use the unadjusted, average normal vowel durations as targets.

There are two reasons why we used context-independent vowel-specific targets as compared to earlier work, which used parallel recordings from a non-dysarthric speaker (Kain et al., 2004). First, there may be a formant mismatch between the targets of the dysarthric and the non-dysarthric speaker. (In our corpus, we had available a female dysarthric speaker and a male non-dysarthric speaker.) For example, a certain CVC combination with female consonants and a male vowel may lead to an abnormal F2 trajectory, not due to a discontinuity, but because of the relative differences between the consonant and vowel formant targets. However, even same-gender speakers may

not be well suited to each other. Second, we hypothesize that it is *not* necessary to map stable-point variance (which may be due to lack of articulatory control or due to formant undershoot) within a vowel, but instead it is only necessary to predict the target directly. For the case of variance due to lack of control, it is *beneficial* to map to the target values; in the case of undershoot, it is *acceptable* to map to the target values. We hypothesize that preserving undershoot throughout the mapping does not aid in improving intelligibility. Furthermore, it then also becomes possible to specify vowel-target values directly without requiring several vowels in various CVC contexts to measure the stable-point variance.

3.4.2. Training

The relationship between dysarthric speech features \mathbf{x} and normal, or intelligible, speech features \mathbf{y} varies depending upon the speakers and is not known *a priori*. Therefore, we had to train a transformation function \mathcal{F} by first establishing a mathematical model of the relationship between \mathbf{x} and \mathbf{y} , and then estimating the parameters θ of this model from a training dataset such that $\mathcal{F}(\mathbf{x}|\theta)$ is a good predictor of \mathbf{y} .

The transformation function was implemented as a mixture of target vectors weighted by posterior probabilities of a Gaussian Mixture Model (GMM). Specifically,

$$\hat{\mathbf{y}} = \mathcal{F}(\mathbf{x}|\theta) = \sum_{q=1}^Q \mathbf{t}_q \cdot p(c_q|\mathbf{x}, \theta) \quad (1)$$

where \mathbf{t}_q is the q th target feature vector from a set of Q target vectors and the term $p(c_q|\mathbf{x}, \theta)$ denotes the GMM posterior probability that the input vector \mathbf{x} belongs to class c_q , given by

$$p(c_q|\mathbf{x}, \theta) = \frac{\alpha_q \cdot \mathcal{N}(\mathbf{x}, \mu_q, \Sigma_q)}{\sum_{i=1}^Q \alpha_i \cdot \mathcal{N}(\mathbf{x}, \mu_i, \Sigma_i)} \quad (2)$$

with

$$\mathcal{N}(\mathbf{x}, \mu, \Sigma) = \frac{e^{-0.5(\mathbf{x}-\mu)\Sigma^{-1}(\mathbf{x}-\mu)}}{(2\pi)^{N/2} \sqrt{\det(\Sigma)}} \quad (3)$$

Table 3
Output target feature values

Vowel (word)	F1 (Hz)		F2 (Hz)		F3 (Hz)		Duration (ms)
	Generic	Individual	Generic	Individual	Generic	Individual	
/i:/ (he)	310	300	2790	2300	3310	2900	212
/I/ (hit)	430	400	2480	1900	3070	2650	138
/E/ (heck)	610	600	2330	1850	2990	2750	167
/@/ (hack)	860	750	2050	1800	2850	2850	257
/u/ (who)	370	350	950	1150	2670	2400	179
/U/ (hook)	470	500	1160	1100	2680	2700	120
/^/ (huff)	760	700	1400	1500	2780	2800	150
/A/ (ha)	850	750	1220	1300	2810	2750	224

Formant frequencies are available as *generic* and *individual* values. The former were taken from Peterson and Barney's work (Peterson and Barney, 1952), while the latter were based on formant frequency values estimated from utterances in the speech corpus.

Model parameters $\theta = (\alpha_{1..Q}, \mu_{1..Q}, \Sigma_{1..Q})$ were estimated by supervised training (full covariances were used). Each vowel was assigned exactly one component ($Q = 8$). Additional experiments showed that using more components led to deteriorating performance, due to the limited number of data points. Formant frequencies were converted from Hertz to the Bark scale, and duration values were given in milliseconds.

An alternative to Eq. (1) would be to simply predict the target vector with the maximum posterior probability. However, unlike in speech recognition, we are not required to reduce acoustic input features to single symbols. Hence, we preferred to “pass on” any ambiguity in the classification of the input dysarthric speech to the transformed speech, in order to maximize chances of the human listener using additional cues to disambiguate the dysarthric speaker’s intent.

Eq. (1) is a variation of the transformation function used in previous work (Kain et al., 2004). Previously, we had modeled the covariance of the joint normal and dysarthric speech features, resulting in the ability to use a mixture of linear transformations for prediction. In the current approach, however, we modeled only the source covariances, leading to a drastic reduction in the number of model parameters. This change was made necessary by our decision to use vowel-specific targets for which covariance data were not available, instead of modeling the distribution of the non-dysarthric speaker in the speech corpus. (As mentioned in Section 3.4.1, we hypothesize that the non-dysarthric speaker’s covariance does not need to be modeled for the goal of improving vowel intelligibility.)

Another consequence of the proposed approach is that it cannot map coarticulatory patterns, as compared to the previous approach. However, we found only weak correlations between coarticulatory patterns in the normal feature space and coarticulatory patterns in the dysarthric feature space, for individual vowels. Moreover, it is unclear whether *mapping* coarticulation is necessary for the goal of improved vowel intelligibility.

We produced models for all 10 rotations and 21 input feature configurations of the training data. An example mapping for configuration 8, displaying F1 and F2 stable points, is shown in Fig. 3.

3.4.3. Objective evaluation

Using the models from the previous section, we classified a new input vector by letting the answer be equal to the class that produces the maximum posterior probability. We defined as a score the number of times that the correct vowel was recognized in this way, normalized by the number of samples in the test set.

Fig. 4 shows the values of these scores, averaged by rotations and by configurations. Given our specific corpus, the optimal input feature set was configuration 8, consisting of the stable points for F1, F2, and F3, together with vowel duration. Generally, the stable-point algorithm outperformed the simple median (configurations 1, 3, 5, and 7 versus configurations 2, 4, 6, and 8). Configuration 8 had the best performance, with an average score of 0.62. Configuration 21 had the second-best average score of 0.58. Neither the addition of F1 nor F2 slopes (configurations 9–12) improved performance over configuration 8. The addition of distance-to-template values (configurations

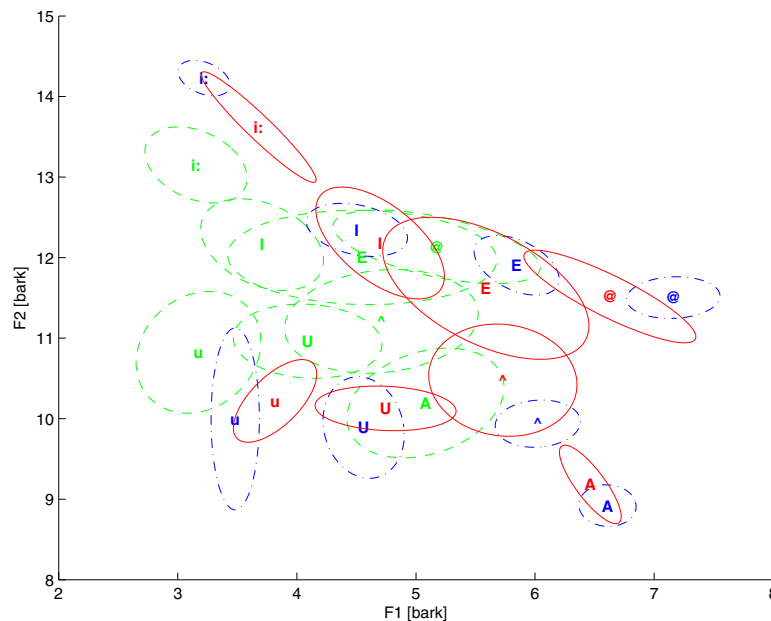


Fig. 3. Bark-scaled F1 and F2 formant frequency stable points for the 8 vowel classes of the training set with ellipses representing Gaussian components (color online). Ellipses are centered at component means and radii are set to one standard deviation. Dashed ellipses represent source (dysarthric) data, dash-dotted ellipses represent target (normal) data, and solid ellipses represent source data after transformation (mapped data).

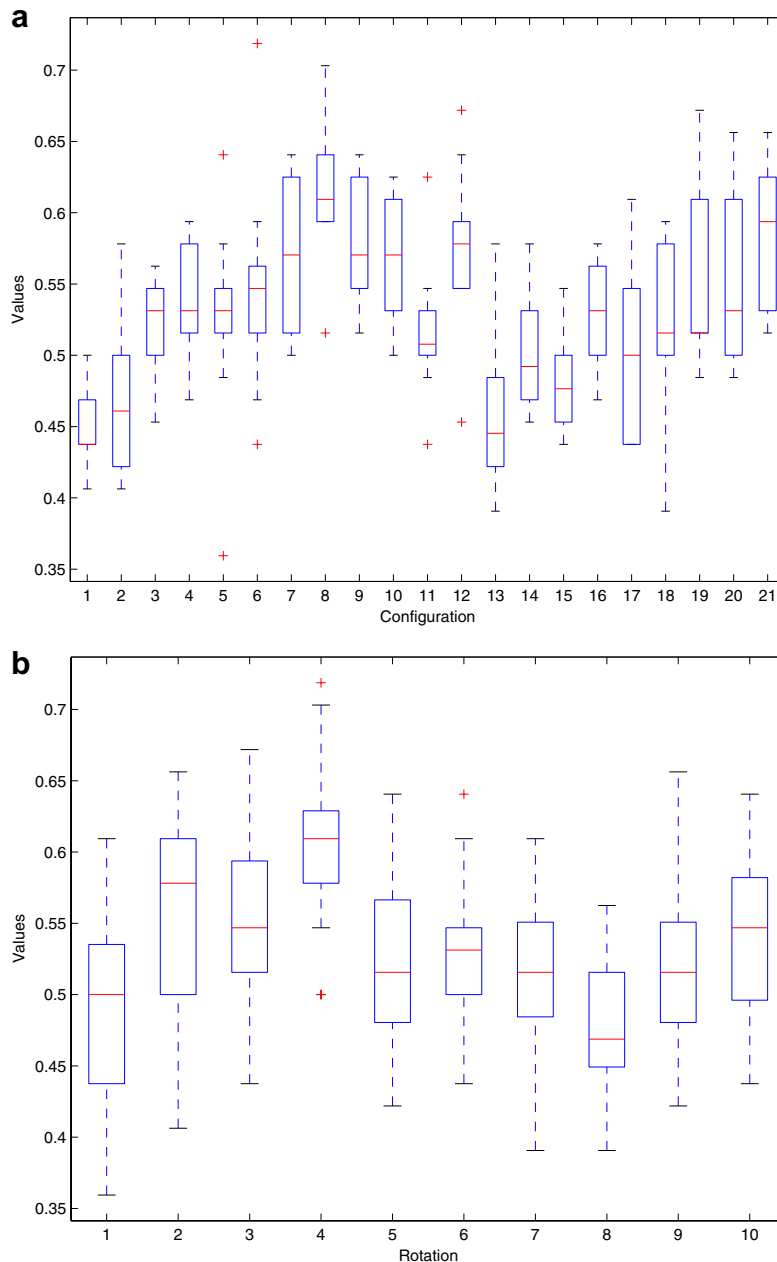


Fig. 4. Boxplot of scores using 21 input feature set configurations and 10 test set rotations. The boxes have lines at the lower quartile, median, and upper quartile values. The whiskers are lines extending from each end of the boxes to show the extent of the rest of the data. Outliers are data with values beyond the ends of the whiskers. (a) Scores averaged over rotations and (b) scores averaged of configurations.

13–16) or polynomial values (configurations 17–20) also did not improve performance. These results may be dependent upon the particular corpus used; adding contextual features may work well with a larger database.

For the perceptual test, we chose the model that was trained on the rotation which had closest to average performance, averaged over all configurations. The model with average performance was chosen so that the results of the perceptual test would show the least bias to both model and test stimuli. Therefore, we used the model trained on configuration 8 and rotation 6 for the remaining sections.

3.5. Synthesis

3.5.1. Feature modification and generation

To obtain the transformed formant trajectory of a dysarthric vowel, we first calculated the stable-point vector of that vowel, and then applied the transformation function to that vector, obtaining the transformed formant targets. In earlier work (Kain et al., 2004), we used a crossfade approach, in which we faded in and out of the original dysarthric trajectory and the formant targets, respectively, to smoothly reach the constant transformed formant targets in the center of the vowel. The crossfade strategy avoided

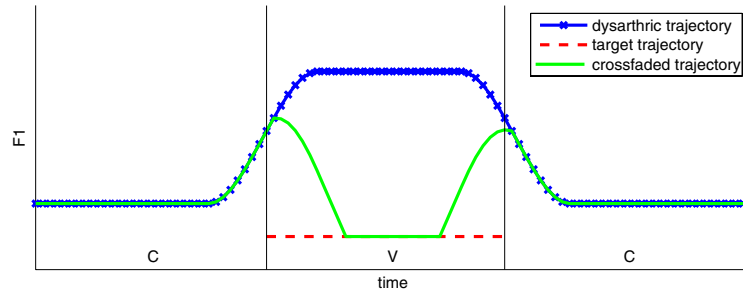


Fig. 5. Example illustrating the use of the crossfade strategy versus a straight-line strategy, using a single hypothetical formant trajectory in a consonant-vowel-consonant context. Even though the crossfade strategy (solid line) prevents discontinuities, it introduces unwanted formant movement, which can lead to incorrect vowel identification. The straight-line strategy, equivalent to the target trajectory (dashed line) at all times throughout the vowel, prevents such movement; its drawbacks are potential discontinuities at the vowel boundaries.

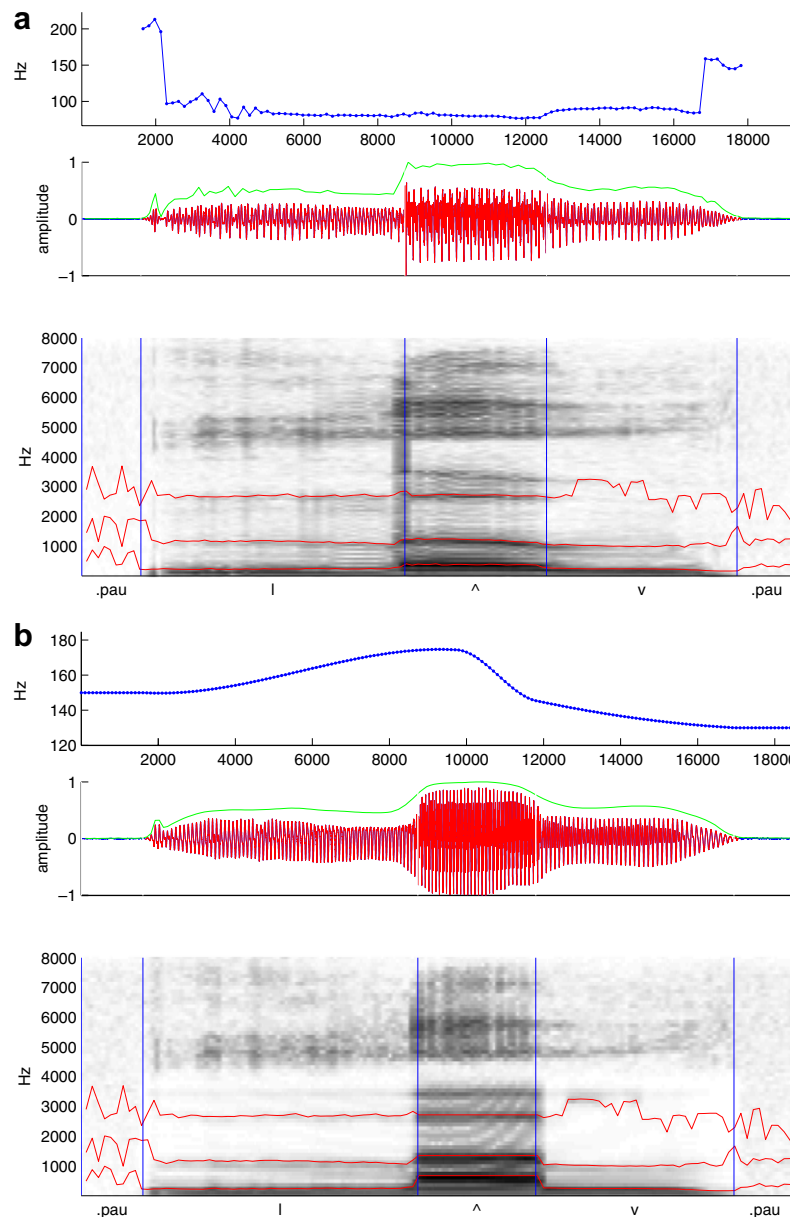


Fig. 6. Analysis of dysarthric speech (a) and synthesis of transformed speech (b). The top panels of each subfigure display F0 values. The middle panels display the speech waveform and energy contour. The bottom panels show spectrograms superimposed with formant frequencies F1–F3.

of this speaker on these material; intelligibility close to 100% was expected.

The *normal-synth-contextdependent* condition (condition G) was an analysis and re-synthesis of the CVC from the non-dysarthric speaker, with formant frequency values based on the vowel in its particular CVC context. (Re-synthesis involved a number of simplifications from the original recording, including synthetic pitch-contour generation, synthetic energy-contour generation, and formant synthesis using a flat formant trajectory throughout the vowel.) Intelligibility somewhat worse than condition H was expected. The *normal-synth-individual* condition (condition F) was an analysis and re-synthesis of the CVC from the non-dysarthric speaker. In this case, formant frequency values were set to those of the individual speaker, but were based only on the vowel identity, not on the vowel's context. The formant frequency values for each vowel in this condition were computed using the interactive procedure described in Section 3.4.1. We expected intelligibility to be somewhat worse than in condition G.

The *dysarthric-oracle-individual* condition (condition E) was an analysis and re-synthesis of the CVC from the dysarthric speaker. In this case, the formant frequency values used in synthesis were the mapping function's *target* values (the normal speaker's formants for the given vowel), representing the outcome of a hypothetical perfect mapping. (The term "oracle" indicates a mapping function that has access to knowledge that would not be present in a normal test condition.) Similar to condition F, the target formant frequency values were the *individual* formant frequency values. We expected intelligibility to be much better than condition A (as a result of using formant frequencies known to yield perception of the correct vowel in isolation), but worse than condition F (because consonant regions were synthetic copies of the dysarthric speech, and vowel identity may depend in part on information in the consonant regions). The *dysarthric-oracle-generic* condition (condition D) was identical with condition E, except that the target formant frequencies were the *generic* values, as described in Section 3.4.2. We expected intelligibility to be close to that of condition E.

The *dysarthric-map-individual* condition (condition C) was an analysis and re-synthesis of the CVC from the dysarthric speaker, but using the results from the mapping function (vowel identity was not used to compute output formant frequencies), and with the individual formant frequencies as targets learned during mapping. This condition represented a true test of the ability to improve the intelligibility of dysarthric speech. We expected intelligibility to be better than intelligibility of condition A (due to the mapping of the vowel space), but worse than condition E (because the mapping function does not always yield the ideal target values on test data). Finally, the *dysarthric-map-generic* condition (condition B) was identical with condition C, except that the target formant values learned by the mapping function were the generic formants. This condition was also a true test of the ability to improve

intelligibility. We expected the intelligibility level to be close to that of condition C, but because we could not safely predict whether B or C should yield better results, both conditions were used in testing.

The comparison of conditions A and B, and the comparison of conditions A and C, tests the effectiveness of the proposed transformation method. Comparing conditions B and D, and comparing conditions C and E, tests the effectiveness of the mapping function. (The mapping function was not expected to be perfect, due to one-to-many mappings inherent in expanding the vowel space, the limited amount of training data, and simple (and probably incomplete) feature representations.) Comparing conditions B and C, and comparing conditions D and E, tests the utility of speaker-independent formant targets against speaker-dependent formant targets. Comparing conditions E and F tests the contribution of consonant regions to vowel identity. Comparing conditions F and G tests the importance of maintaining context-dependent units in this (flat-trajectory) synthesis framework, and comparing conditions G and H tests the quality of the flat-trajectory formant synthesis, pitch generation, and energy generation.

4.2. Test administration

With 64 CVCs in eight conditions, a total of 512 unique stimuli were available for evaluation. Evaluation was conducted by 24 listeners, each of whom evaluated 128 stimuli. Each of the 512 stimuli were evaluated six times, by a different listener each time. Each listener heard all CVCs twice, with the two presentations of a CVC in different conditions each time. (Subsequent analysis of results showed that performance on a CVC did not, on average, improve on the second presentation of that CVC, and so learning effects were minimal.) The order of presentation of the CVCs was randomized, but all listeners heard the CVCs in the same (random) ordering, to normalize the effect of the order of CVC presentation on different listeners.

We developed a perceptual test for vowels based on the procedures described by Kent et al. (1989). Listeners heard a stimulus and indicated the identity of the vowel that they perceived. The intelligibility was computed as the number of correctly perceived vowels divided by the total number of vowels. Listening tests were conducted over speakers in a quiet, but not sound-isolated, room, in order to better match the test conditions with an expected real-world condition of transformed speech played over speakers. Listeners reported normal hearing, were native speakers of American English, and had no clinical or research experience with dysarthria. Listeners were paid for their participation.

Testing was conducted using a graphical user interface designed for this experiment. After reading instructions presented on a screen and entering a listener ID number, the listener was guided through three familiarization stages. In the first stage, a fixed set of eight words were displayed on the screen. Each single-syllable word was

pronounced with one of the eight vowels used in this study, namely “heed” (for /i:/), “hid” (for /I/), “heck” (for /E/), “had” (for /@/), “who” (for /u/), “hook” (for /U/), “hut” (for /^/), and “hot” (for /A/). Clicking on a word caused the vowel associated with that word to be played over the speakers. Listeners were encouraged to familiarize themselves with the words and associated vowel sounds, and each word had to be clicked on at least once before proceeding to the second stage. In the second stage, the listener went through a process similar to the actual test. The listener heard a CVC “word,” and had to select the word on the screen with the vowel most similar to the word that they heard. At this stage, all words were spoken by the non-dysarthric speaker, JH. This stage familiarized the listener with the format of the test, while providing relatively easy tasks. After identifying 10 of these stimuli, the user proceeded to the third stage of familiarization. The third stage was identical to the second stage, except that there were six samples from the dysarthric speaker, DC. One of these samples was simply a CVC recording; other samples were synthetic CVCs similar to the stimuli in the test. This stage familiarized the listener with the type of stimuli that would be heard during the actual test. After completing the three familiarization stages, the listener evaluated the 128 test stimuli. Recorded information included the correct (intended) vowel identity, the perceived vowel identity, the stimulus condition, and the time required to provide a response. (The time information, however, was not used in the final evaluation.)

Based on our previous perceptual tests using these CVC stimuli, we expected intelligibility of the normal condition to be close to 100%. If a listener scored dramatically lower than expected, this was considered evidence that the listener either had an unreported hearing problem, or that the listener had difficulty performing the conceptual mapping from heard CVC (e.g. “push”) to the word on the screen with the most similar vowel (e.g. “hook”). Neither of these conditions were considered relevant to the goal of evaluating vowel intelligibility, and so we set a threshold of at least 90% reported intelligibility on the normal condition for each listener. We conducted testing until 24 listeners passed this threshold of acceptance, eliminating 9 listeners in the process.

In addition, we conducted this test on a person who was extremely familiar with dysarthric speech in general and also familiar with the speech of the current dysarthric speaker, DC. We would like, as a long-term target for the performance level of a transformation system, to do as well with an automatic system as an expert human. While it is theoretically possible to have better-than-human performance by a transformation system (given the nature of human variance), we consider for now human-expert levels of performance to be the practical limit of technology. In addition, the test results of the expert listener may yield interesting insights into the ability of the transformation system to improve intelligibility for expert listeners.

4.3. Results and discussion

Results of the perceptual test averaged over the 24 listeners are shown in Table 5. The results of a planned statistical comparison test showed that both the dysarthric-map-generic and dysarthric-map-individual conditions performed significantly better (from 48% to 54%, a 6% improvement) than the natural dysarthric speech ($p = 0.017$ and $p = 0.030$, respectively); however, there was no statistical difference between the two mapping conditions. The effect size between the dysarthric and the dysarthric-map-individual conditions, using the pooled standard deviation, was $d = 0.53$, considered “medium” (Cohen, 1988).

In addition, 16 out of 24 subjects had better results on the average of dysarthric-oracle-generic and dysarthric-oracle-individual conditions compared with the dysarthric condition, indicating that most listeners had better performance on the mapped speech than the original speech.

Analyzing individual vowel performance, we observe that vowel height is strongly correlated to the intelligibility of the original dysarthric speech. The mapping function improves the intelligibility of the vowels with the lowest height (/@/ and /A/) most effectively, improves the intelligibility of the next-lowest vowels (/E/ and /^/) less effectively, while proving detrimental to the original dysarthric intelligibility of the high vowels (/i:/, /u/, and /U/). It is possible that the algorithm performed best for

Table 5
Intelligibility of stimulus conditions in percent

Stimulus condition	/i:/	/I/	/E/	/@/	/u/	/U/	/^/	/A/	Average	Expert
A – dysarthric	73	63	40	10	92	73	27	6	48 (13)	69
B – dysarthric-map-generic	67	42	54	83	46	52	19	73	54 (10)	56
C – dysarthric-map-individual	50	65	56	83	56	56	21	46	54 (10)	75
D – dysarthric-oracle-generic	96	42	77	94	81	63	71	92	77 (11)	81
E – dysarthric-oracle-individual	94	83	88	96	88	63	54	71	79 (9)	100
F – normal-synth-individual	96	88	92	98	98	73	96	94	92 (8)	88
G – normal-synth-contextdependent	96	83	83	98	71	79	94	98	88 (9)	100
H – normal	100	98	100	100	98	92	100	100	98 (3)	100

Values are for 24 non-expert listeners, except for the last column which contains results from 1 expert listener. Standard deviations are given in parentheses in the penultimate column.

low vowels because the separation between F0 and F1 is maximal; another possibility is that performance was best for low vowels because of the greater overlap between vowels along the F1 dimension, compared with the F2 dimension, for the dysarthric speaker. If one could detect and exclude high vowels from processing and only transform medial and low vowels, the rate of intelligibility would increase to about 63%, for a 15% improvement.

Tables 6(a) through (d) show confusion matrices for the dysarthric, dysarthric-map-individual, dysarthric-oracle-individual, and normal-synth-individual conditions. Studying the first confusion matrix, we observe that most errors are made in judging the height and the duration of the vowel, for example responding /I/ for a given /E/ and responding /^/ for a given /A/, respectively. Typically, listeners judged the height of the vowel as too high. Less confusion occurred with regard to the backness of the vowel. The next confusion matrix in Table 6 reflects the changes that occurred after applying the transformation to the

dysarthric speech. We observe that the front vowels are now less likely to be confused with the back vowels, and vice versa, with the exception of the vowel /^/. Additionally, the mapping “adjusted” the height of the vowels so that listeners are now less likely to judge vowels as too high, at the expense of judging vowels too low more often than before. Similar observations can be made in the case of the dysarthric-oracle-individual and normal-synth-individual condition results displayed in Tables 6(c) and (d).

Most of our expectations of relative intelligibility (Section 4.1) were met. In particular, dysarthria-map-generic and dysarthria-map-individual were both better than the dysarthria condition, the dysarthria-oracle conditions had greater intelligibility than the dysarthria-map conditions, the dysarthria-map conditions had lower intelligibility than the normal-synth conditions, and the normal-synth conditions had lower intelligibility than the normal condition. One exception to our expected performance was that the normal-synth-individual condition had greater intelligibility than the normal-synth-contextdependent condition; however, the difference in intelligibility was not statistically significant.

In summary, testing showed that the proposed mapping method is statistically significantly more intelligible than the original dysarthric speech for this dysarthric speaker. However, the transformation function does not yield nearly the same intelligibility levels as the oracle condition, indicating that the mapping is quite imperfect. The use of speaker-independent targets or speaker-dependent targets had, for the average listener, no effect on intelligibility. The large difference in intelligibility between the dysarthria-oracle conditions and the normal-synth-individual conditions indicates that a large amount of information about the vowel identity is located in the consonant regions of the CVCs, as the only difference between these conditions was in those regions. This is consistent with many previous studies on the importance of consonants on vowel classification (Strange et al., 1976). The use of context-dependent vowel targets had no significant impact, compared with context-independent vowel targets. The synthesis framework, including the flat-trajectory model and other simplifications, caused a dramatic decrease in intelligibility, from 98% to approximately 90%.

Conditions B–E represent changes to a number of acoustic features simultaneously, including F0, vowel duration, source characteristics, formant frequencies, and formant bandwidths. Therefore, our results report on the combined effect only; the individual contribution of each modification to improved intelligibility is unknown.

The results for the single expert listener illustrate a number of interesting points. First, the proposed method does not yield intelligibility results for the average listener that are as good as intelligibility of the original speech heard by the expert listener. However, the transformation system, when used with the individual’s formant targets, did improve intelligibility even for the expert listener (from 69% to 75%). When the transformation system was used with

Table 6

Confusion matrix for selected conditions. Target and response vowels are given in the first column and the first row, respectively

Vowel	/i:/	/I/	/E/	/@/	/u/	/U/	/^/	/A/
<i>(a) Condition A – dysarthric</i>								
/i:/	35	7	2	0	2	2	0	0
/I/	2	30	4	0	7	4	1	0
/E/	0	18	19	1	1	4	5	0
/@/	0	13	22	5	0	4	4	0
/u/	0	0	0	0	44	3	1	0
/U/	0	2	0	0	10	35	1	0
/^/	0	8	7	0	7	11	13	2
/A/	0	1	0	1	1	14	28	3
<i>(b) Condition C – dysarthric-map-individual</i>								
/i:/	24	13	2	1	1	6	0	0
/I/	7	31	8	0	2	0	0	0
/E/	0	8	27	11	0	0	2	0
/@/	0	0	8	40	0	0	0	0
/u/	0	0	0	0	27	15	6	0
/U/	0	0	1	2	7	27	11	0
/^/	0	13	15	3	0	0	10	7
/A/	0	0	0	3	0	3	20	22
<i>(c) Condition E – dysarthric-oracle-individual</i>								
/i:/	45	2	0	0	1	0	0	0
/I/	2	40	3	0	0	2	1	0
/E/	0	1	42	3	0	0	2	0
/@/	0	0	2	46	0	0	0	0
/u/	0	0	0	0	42	6	0	0
/U/	0	0	0	0	1	30	17	0
/^/	0	0	6	16	0	0	26	0
/A/	0	0	0	7	0	0	7	34
<i>(d) Condition F – normal-synth-individual</i>								
/i:/	46	2	0	0	0	0	0	0
/I/	3	42	3	0	0	0	0	0
/E/	0	0	44	4	0	0	0	0
/@/	0	0	0	47	1	0	0	0
/u/	0	0	0	0	47	1	0	0
/U/	0	0	0	0	3	35	10	0
/^/	0	0	0	0	0	0	46	2
/A/	0	0	0	2	0	0	1	45

generic formant targets, however, intelligibility for this expert listener decreased below the intelligibility level of the original speech. It is unclear whether this difference in performance from individual targets and generic targets is due to familiarity of the expert listener with the dysarthric speaker's voice, larger concatenation errors between vowel and consonant for the generic formant targets, or some other factor(s). A similar pattern is seen in the oracle condition; the generic formant targets yield intelligibility of 81%, while the individual formant targets yield vowel intelligibility of 100%, which is even greater than the intelligibility of the normal-synth-individual condition (88%). This particular expert listener appears to utilize context-dependent vowel information, as her normal-synth-individual result was 88%, but her normal-synth-contextdependent result was 100%. In general, this listener seems to listen to the *entire* CVC in order to determine the vowel identity, even more so than the average listener. It is unfortunate that more such expert listeners were not available to us, which prevents any claims of statistical significance from being made.

5. Conclusion

In this study, we have significantly improved the intelligibility of dysarthric vowels from 48% to 54%, as evaluated by a vowel identification task using 64 CVC stimuli judged by 24 listeners. The optimal mapping feature set from a list of 21-candidate feature sets proved to be one utilizing vowel duration and F1–F3 stable points, which were calculated using shape-constrained isotonic regression. The choice of speaker-specific or speaker-independent vowel formant targets appeared to be insignificant. Weaknesses in the transformation function were obviated through comparisons with “oracle” conditions. Also, the synthesis-framework itself has been shown to have a negative impact on intelligibility. The authors would like to stress that the results in this paper are highly preliminary, as they pertain only to a single person with dysarthria and cannot be generalized to dysarthric speech in general. Even when categorized as having the same type of dysarthria, speakers are enormously variable in their speech characteristics. Other speakers may have worse or better levels of improvement. Moreover, the effect size was merely medium.

To extend our work to sentence-level processing, additional challenges must be met. Firstly, a consonant-vowel boundary detector must be implemented that yields satisfactory results on dysarthric speech. Secondly, additional consideration must be given to diphthongs, which consist of more than a single stable point. Finally, F0 prediction will be more complex, and potentially semantically meaningless, although the dysarthric speech energy trajectory may be used for cues about word emphasis and syllable stress. Even with these additions, such an algorithm would be relatively “light-weight” and executable in real-time on a wearable computer.

Possible avenues to further increase intelligibility include:

- Instead of using formant stable points as transformation features, the mapping function may benefit from using formant frequency values that have been “de-coarticulated” from their surrounding environment, using an approach similar to that taken by Niu and van Santen (2003).
- It has been shown that even just two formant frequency targets (at 20% and 80%) perform better than flat formant frequency trajectories (Hillenbrand and Nearey, 1999). Therefore, a more sophisticated formant trajectory model is likely to improve the intelligibility of dysarthric speech.
- In the case of using generic formant target values, it may be better to choose a naturally large vowel space (Bradlow et al., 1996) to maximize the acoustic distance between vowels.
- In previous experiments (Hosom et al., 2003), we have shown that consonants contribute greatly to improved overall intelligibility. Moreover, consonants also contribute specifically to vowel intelligibility, as demonstrated by our current results. Therefore, work on transforming consonants is likely to bring improvements to overall intelligibility. However, variability in dysarthric consonants forms a significant hurdle.
- Additional non-acoustic inputs describing the speech production state may aid in transformation accuracy. We are currently experimenting with a non-invasive device that can give reliable information about the status of the velopharyngeal port (Niu et al., 2006).
- A highly intelligible, but much more error-prone approach consists of coupling a speech recognizer with a Text-to-Speech system, functioning as a “language-interpreter” device. Some research studying dysarthric speech recognition has already been carried out (Ferrier et al., 1995).

Acknowledgements

This research was conducted with support from NSF Grant 0117911 “Making Dysarthric Speech Intelligible”. Oregon Health & Science University (OHSU), Dr. Kain, Dr. Hosom, and Dr. Jan van Santen have a significant financial interest in BioSpeech, Inc., a company that may have a commercial interest in the results of this research and technology. This potential conflict was reviewed and a management plan approved by the OHSU Conflict of Interest in Research Committee and the Integrity Program Oversight Council was implemented.

References

- Barlow, R.E., Barholomew, D.J., Bremner, J.M., Brunk, H.D., 1980. Statistical Inference under Order Restrictions. Wiley, Chichester.

- Beukelman, D.R., Mirenda, P., 2005. *Augmentative and Alternative Communication Management of Severe Communication Disorders in Children and Adults*, third ed. Paul H. Brookes, Baltimore, MD.
- Bradlow, A.R., Toretta, G.M., Pisoni, D.B., 1996. Intelligibility of normal speech: Global and fine-grained acoustic–phonetic talker characteristics. *Speech Comm.* 20, 255–272.
- Carnegie Mellon University, 2004. The CMU Pronouncing Dictionary v0.6. <<http://www.speech.cs.cmu.edu>>.
- Cohen, J., 1988. *Statistical Power Analysis for the Behavioral Sciences*, second ed. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Darley, F., Aronson, A., Brown, J., 1969. Differential diagnostic patterns of dysarthria. *J. Speech Hearing Res.* 12 (2), 246–269.
- Di Benedetto, M.-G., 1989. Vowel representation: Some observations on temporal and spectral properties of the first formant frequency. *J. Acoust. Soc. Amer.* 86 (1), 55–66.
- Drager, K., Hustad, K., Gable, K., 2004. Telephone communication: Synthetic and dysarthric speech intelligibility and listener preferences. *Augment. Alternat. Comm.* 20 (2), 103–112.
- Duffy, J., 2005. *Motor Speech Disorders: Substrates, Differential Diagnosis and Management*, second ed. Mosby, St. Louis, MO.
- Electronic Speech Enhancement Inc., The Speech Enhancer. <<http://www.speechenhancer.com>>.
- Ferrier, L., Shane, H., Ballard, H., Carpenter, T., Benoit, A., 1995. Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. *Augment. Alternat. Comm.* 11 (3), 165–175.
- Hartelius, L., Runmarker, B., Andersen, O., 2000. Prevalence and characteristics of dysarthria in a multiple-sclerosis incidence cohort: Relation to neurological data. *Int. J. Phoniatrics Speech Therapy Comm. Pathol.* 52 (4), 160–177.
- Hieronymus, J.L., ASCII Phonetic Symbols for the World's Languages: Worldbet, Bell Labs Technical Memorandum.
- Hillenbrand, J.M., Nearey, T.M., 1999. Identification of resynthesized /hVd/ utterances: Effects of formant contour. *J. Acoust. Soc. Amer.* 105 (6), 3509–3523.
- Hosom, J.P., Kain, A.B., Mishra, T., van Santen, J.P.H., Fried-Oken, M., Staehely, J., 2003. Intelligibility of modifications to dysarthric speech. In: *Proc. of ICASSP*, pp. 878–881.
- Kain, A.B., Niu, X., Hosom, J., Miao, J., van Santen, 2004. Formant resynthesis of dysarthric speech. In: *IEEE Workshop on Speech Synthesis*, Pittsburgh, PA, pp. 25–30.
- Kent, R.D., Weismer, G., Kent, J.F., Rosenbek, J.C., 1989. Toward phonetic intelligibility testing in dysarthria. *J. Speech Hearing Disorders* 54, 482–499.
- Kent, R.D., Kent, J.F., Weismer, G., Sufit, R.L., Rosenbek, J.C., Martin, R.E., Brooks, B.R., 1990. Impairment of speech intelligibility in men with amyotrophic lateral sclerosis. *J. Speech Hearing Disorders* 55 (4), 721–728.
- Klatt, D., 1987. Review of text-to-speech conversion for English. *J. Acoust. Soc. Amer.* 82 (3), 737–793.
- Lindblom, B., 1963. Spectrographic study of vowel reduction. *J. Acoust. Soc. Amer.* 35 (11), 1773–1781.
- Maassen, B., Povel, D.J., 1984. The effect of correcting fundamental frequency on the intelligibility of deaf speech and its interaction with temporal aspects. *J. Acoust. Soc. Amer.* 76 (6), 1673–1681.
- Maassen, B., Povel, D.J., 1985. The effect of segmental and suprasegmental corrections on the intelligibility of deaf speech. *J. Acoust. Soc. Amer.* 78 (3), 877–886.
- Menéndez-Pidal, X., Polikoff, J.B., Peters, S.M., Leonzio, J.E., Bunnell, H., 1996. The nemours database of dysarthric speech. In: *Proc. of ICSLP*, Philadelphia, PA, vol. 3, pp. 1962–1965.
- Niu, X., van Santen, J.P.H., 2003. 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 Internat. Workshop, pp. 233–236.
- Peterson, G.E., Barney, H.L., 1952. Control methods used in a study of the vowels. *J. Acoust. Soc. Amer.* 24 (2), 175–184.
- Prentke Romich Company, The Pathfinder Communications Device. <<http://www.prentrom.com>>.
- Qi, Y., Weinberg, B., Bi, N., 1995. Enhancement of female esophageal and tracheoesophageal speech. *J. Acoust. Soc. Amer.* 98 (5), 2461–2465.
- Semantic Compaction Systems, The Minspeak Language Representation Technique. <<http://www.minspeak.com>>.
- Shuster, L.I., 1996. Linear predictive coding parameter manipulation/synthesis of incorrectly produced /t/. *J. Speech Hearing Res.* 39 (4), 827–832.
- Stevens, K.N., House, A.S., 1963. Perturbation of vowel articulations by consonantal context: An acoustical study. *J. Speech Hearing Res.* 6 (2), 111–128.
- Strange, W., Shankweiler, D.P., Edman, T.R., 1976. Consonant environment specifies vowel identity. *J. Acoust. Soc. Amer.* 60 (1), 213–224.
- Titze, I., 2000. *Principles of Voice Production*, National Center for Voice and Speech, Iowa City, IA.
- Turner, G.S., Tjaden, K., Weismer, G., 1995. The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis. *J. Speech Hearing Res.* 38, 1001–1013.
- van Santen, J.P.H., Möbius, B., 2000. A quantitative model of F0 generation and alignment. *Intonation – Analysis, Modelling and Technology*. Kluwer, Dordrecht, pp. 269–288.
- Visser, J., pVoice: Dynamic Screen Communication Software. <<http://www.pvoice.org>>.
- Weismer, G., Yunusova, Y., Westbury, J.R., 2003. Interarticulator coordination in dysarthria: An X-ray microbeam study. *J. Speech Lang. Hearing Res.* 46 (5), 1247–1261.
- Words+ Inc., “Say-it! SAM” tablet computer including software titles “E Z Keys”, “Speaking Dynamically Pro”, “Boardmaker, and Talking Screen”. <<http://www.words-plus.com>>.
- X. Niu, A.B. Kain, J.P. van Santen, 2006. A noninvasive, low-cost device to study the velopharyngeal port during speech and some preliminary results. In: *Proc. of ICSLP*, pp. 957–960.
- Yorkston, K.M., Beukelman, D.R., Bell, K.R., 1988. *Clinical Management of Dysarthric Speakers*. PRO-ED, Inc., Austin, TX.
- Yorkston, K., Beukelman, D., Strand, E., Bell, K., 1999. *Management of Motor Speech Disorders in Children and Adults*. PRO-ED, Inc., Austin, TX.
- Ziegler, W., Hartmann, E., von Cramon, D., 1988. Word identification testing in the diagnostic evaluation of dysarthric speech. *Clin. Linguistics Phonetics* 2, 291–308.