Abstract: Dysarthria is a diverse group of motor speech disorders that typically are associated with impaired intelligibility. As part of a project to develop augmentation communication technologies for intelligibility enhancement of dysarthric speech, a quantitative method is proposed for measuring the relative contributions to impaired intelligibility of vowels of three factors: First, target shift: Dysarthric speakers may have spectral targets that differ from those of normal speakers. Second, coarticulation: The degree of contextual influence on articulation may be greater in dysarthric speech than in normal speech. Third, random variability: Dysarthric speakers may articulate the same phoneme in the same context with more variability. The method is based on a linear model of formant trajectories of vowels in consonant contexts. The results from analysis of a dysarthric and a normal speech sample showed surprisingly similar target values, but increased coarticulation and random variability for the dysarthric sample.

Keywords: Dysarthria, coarticulation, formant

I. INTRODUCTION

Dysarthria is a diverse group of motor speech disorders that typically are associated with impaired intelligibility and are caused by damage to the motor system [1, 2]. Since in most cases dysarthria is not reversible, major efforts have been made to create assistive devices, including devices based on speech enhancement [3], speech recognition [4], or speech transformation [5].

A recent perceptual study by Hosom et al. [5] focused on the relative contributions of segmental and prosodic factors to intelligibility of dysarthric speech. Using a human-supervised copy prosody technique that allowed for the independent modification of prosodic and spectral information in dysarthric speech, it was shown that significant improvements of intelligibility can be achieved through replacing either the prosodic features or the spectral features of a dysarthric speaker's speech with those of a normal speaker's speech. However, an automated baseline transformation system, based on speech transformation techniques to map the spectral features between the two speakers on a frame-by-frame basis [6], failed to improve intelligibility. A further analysis of the vowel formants indicated that their average values differ sharply between the dysarthric and the normal speech samples, with a much-reduced area of the vowel quadrilateral in the former case [Figs. 1 and 2].

These findings show that successful intelligibility enhancement requires an underlying model of the spectral differences between dysarthric and normal speech. Towards such a model, we consider here three factors that may account for these differences. First, target shift: Dysarthric speakers may develop special spectral targets that differ from those of normal speakers. Second, coarticulation: The degree of contextual influence on articulation may be greater in dysarthric speech than in normal speech. Third, random variability: Dysarthric speakers may articulate the same phoneme in the same context with more variability.

This paper provides an analysis approach that decomposes the contributions of these factors, so that they can be treated separately in the future intelligibility enhancement systems. This analysis will be applied to speech samples from one dysarthric and one normal speaker for demonstration purposes only; no claims are made about dysarthric speech in general.

Since formants constitute a concise acoustic representation closely related to the vocal-tract configuration, we focus our investigation on formant trajectories. We use a linear superposition model similar to a model by Broad and Clermont [7] to describe the trajectories of the first three formants through inter-consonantal vowel portions. In our model, target formants and coarticulatory effects are unknown parameters and are estimated from speech data. Beyond the structure of the model, nothing is assumed about these parameters, so that their estimated values provide unbiased information about the differences between dysarthric and normal speech. The experiments on dysarthric and normal speech data show surprisingly similar target values, but increased coarticulation and random variability for the dysarthric sample. We expect that these results can be used to construct an augmentative communication system.

II. METHODOLOGY

A. Speech data

For the purpose of comparison, the same speech data as in the previous study [5] were used. The data are utterances of one dysarthric speaker (LL) and one normal speaker (JP) from the Nemours database (For diagnostic information, see [8]). Each speaker read 74 syntactically correct nonsense sentences. The speech was recorded and stored in 16k Hz, 16-bit PCM format.
Each sentence in the database has been transcribed into a sequence of phoneme labels. The start and end times of each phoneme in the speech signals were indicated via manual segmentation. The segments considered in this study were syllables consisting of two consonants (CVC) in American English. The vowels consisted of /i/, /u/, /A/, and /@/, as pronounced in the words beat, boot, father, and bad. They are supposed to represent four extreme vocal-tract configurations among the vowels in American English. The consonants consisted of the six stops, the four unvoiced fricatives, and the four approximants in American English.

ESPS software [9] was used to extract formant trajectories from the speech signals. The signals were down-sampled from 16k Hz to 10k Hz, and analyzed with a 49-ms Hanning window that was shifted in a 10-ms step. For each frame of windowed signals, a 12-order LPC analysis was performed and then continuous formant trajectories were obtained. Formant values at vowel midpoints were inspected and, when necessary, corrected manually with optional LPC-poles.

B. Coarticulatory model

We adopted the following model, similar to which was used in [7], to describe coarticulatory effects on the formant frequencies of vowels within different consonant contexts:

\[
\tilde{F}(t) = \alpha(t) \cdot (\tilde{T}_c - \tilde{T}_v) + \tilde{T}_v + \beta(t) \cdot (\tilde{T}_c - \tilde{T}_v),
\]

(1)

where \(\tilde{F}(t)\) is the observed formant vector as a function of time \(t\), \(\tilde{T}_c\) is the target formant-vector of the vowel, \(\tilde{T}_v\) and \(\tilde{T}_c\) are the target formant-vectors of the initial and final consonants, respectively. All formant vectors in Eq. (1) are 3x1 in dimension with the first three formant frequencies as elements. The first term in the right side of Eq. (1) represents formant transitions from the consonant \(C\) to the vowel \(V\). This coarticulatory effect is proportional to the target difference and scaled by a function of the coarticulatory factor \(\alpha(t)\). The last term represents a similar effect of the consonant \(C'\) on the vowel \(V\), and \(\beta(t)\) is the corresponding function of the coarticulatory factor.

If we let \(\gamma(t) = (1 - \alpha(t) - \beta(t))\), then Eq. (1) becomes:

\[
\tilde{F}(t) = \alpha(t) \cdot \tilde{T}_c + \gamma(t) \cdot \tilde{T}_v + \beta(t) \cdot \tilde{T}_c,
\]

(2)

which shows that the observed formant vector of the vowel at any time point is a linear combination of the target formant-vectors of the phonemes \(C, V\) and \(C'\).

The model represents the three factors (target shift, coarticulation, and random variability) as follows: Target shift is represented by the differences in target values between the dysarthric and normal speaker; coarticulation by the values of the coarticulatory factors; and random variability by the relative goodness of fit of the model.

C. Estimation method

\(N\) denotes the number of samples of observed formant vectors, \(\tilde{F}^{(i)} (i = 1, \ldots, N)\). If target formant vectors are known, a least-square-error solution exists for each of the following equations derived from Eq. (1):

\[
\begin{bmatrix}
\tilde{F}^{(1)} - \tilde{F}^{(1)} \\
\tilde{F}^{(2)} - \tilde{F}^{(2)} \\
\vdots \\
\tilde{F}^{(N)} - \tilde{F}^{(N)}
\end{bmatrix}
= 
\begin{bmatrix}
\tilde{T}_c^{(1)} - \tilde{T}_v^{(1)} \\
\tilde{T}_c^{(2)} - \tilde{T}_v^{(2)} \\
\vdots \\
\tilde{T}_c^{(N)} - \tilde{T}_v^{(N)}
\end{bmatrix}
\begin{bmatrix}
\alpha^{(1)} \\
\beta^{(1)} \\
\vdots \\
\beta^{(N)}
\end{bmatrix}
\]

(3)

When \(\alpha^{(i)}\) and \(\beta^{(i)}\) are fixed, Eq. (2) can also be rewritten in the following matrix form:

\[
\tilde{F}^{(i)} = \begin{bmatrix}
\alpha^{(i)} \cdot I - (1 - \alpha^{(i)} - \beta^{(i)}) \cdot I \\
\beta^{(i)} \cdot I
\end{bmatrix}
\begin{bmatrix}
\tilde{T}_c^{(i)} \\
\tilde{T}_v^{(i)}
\end{bmatrix}
\]

(4)

where \(I\) is a 3x3 identical matrix. Since the phonemes with the same identity in the samples share a common target formant-vector, all equations in (4) can be jointly solved in a least-square-error sense as long as the number of data samples is large enough. Thus, the estimation algorithm can be generally described as follows:

1. Initialize target formant-vectors;
2. Set a small number \(\varepsilon\) as the convergence threshold;
3. Solve equations in (3), update \(\alpha^{(i)}\) and \(\beta^{(i)}\), and calculate the square error \(E_1\);
4. Solve equations in (4), update target formant-vectors, and calculate the square error \(E_2\);
5. If \([E_1 - E_2] > \varepsilon\), then go to 3; else, output target formant-vectors, \(\alpha^{(i)}\) and \(\beta^{(i)}\).

D. Practical issues

When using this method to analyze the real speech data, additional efforts are needed to avoid physically meaningless solutions. This is discussed next.

Formant target initialization. One scheme adopted the formant values from the Klatt synthesizer [10]. For the vowels, we also used the medians of observed formants at vowel midpoints. The vowel targets estimated with the two initializing schemes were quite close.
Figure 1. Observed and target vowel formants (dysarthric speech.) The medians of observed formants are linked with dashed lines; the estimated target formants are linked with solid lines.

Rank. When in Eq. \((3) \bar{T}_C^{(i)} = \bar{T}_C'^{(i)}\), the matrix on the right is not full-ranked, so that \(\alpha^{(i)}\) and \(\beta^{(i)}\) cannot be estimated. Thus, \(C\) and \(C'\) are required to be different.

Constraints. Constraints of the linear weights included: \(\alpha^{(i)} \geq 0.025, \beta^{(i)} \geq 0.025\), and \(\alpha^{(i)} + \beta^{(i)} \leq 0.95\). For a target formant vector, \(\bar{T} = [f_1, f_2, f_3]\), constraints were \(90 \leq f_1 \leq 1300, 500 \leq f_2 \leq 2800, 1300 \leq f_3 \leq 3700\), and \(f_1 < f_2 < f_3\), because the formants should be in reasonable ranges.

Normalization. Note that \(\alpha^{(i)}\) and \(\beta^{(i)}\) are scalar values, while formants are vectors. Hence, \(\alpha^{(i)}\) and \(\beta^{(i)}\) reflect only the average coarticulatory effects of the three formant frequencies. To balance the contributions of each formant to the fitting errors, each dimension of a formant vector was normalized by dividing it by the formant medians.

III. RESULTS

A. Goodness of fit

The normalized sums of least squares deviations were 0.238 and 0.032 for the dysarthric and normal samples, respectively, indicating greater variability for the dysarthric speech.

B. Vowel space

Figs. 1 and 2 show the observed vowel space and the estimated target vowel space of the dysarthric and normal speaker, respectively. In both figures, the first and second formants (F1, F2) of four extreme vowels (/i, u, A, @/) are plotted as points in the F1-F2 plane. The medians of observed formants are linked with dashed lines to represent an observed vowel space, and the estimated values of target formants are linked with solid lines to represent a target vowel space. As can be seen, the formant quadrilateral for each speaker shifts from the observed position to the target position, expanding the area of the vowel space. This expansion trend implies a potential way to increase the spectral separability of these vowels, which may be critical for intelligibility enhancement, by considering the target formants rather than the observed formants. Of critical importance is that, except for /@/, the dysarthric target values are surprisingly close to the normal target values.

IV. DISCUSSION AND CONCLUSION

In summary, an approach to formant analysis was presented that decomposes the contributions of target...
formants and coarticulatory effects on the formant trajectories of CVC syllables. The approach adopts a linear superposition model to describe formant trajectories. Using the method, target formants and coarticulatory factors can be estimated from speech data. Using the method, we analyzed the speech data of a dysarthric speaker and a normal speaker to gain insight in the relative contributions of three factors that may be responsible for reduced intelligibility of dysarthric speech: target shift, coarticulation, and random variability. The results from this preliminary experiment revealed systematic differences between the two speakers. The target vowel space of the dysarthric speaker exhibits a specific distortion pattern of vowel production, but was surprisingly similar to the target space of the normal speaker. The analysis results also show a larger degree of coarticulatory effects in the speech of this dysarthric speaker, and more random variability.

The analysis results show that intelligibility enhancement may critically need algorithms for the “decoarticulation” of dysarthric speech. In principle, if the system can recognize aspects of vowel environments, such as the place of articulation of surrounding consonants, this could be accomplished by applying Eq. (2) in reverse to recovered the true vowel formants from the observed formants and the inferred consonant targets.

We note that the model is extremely simple. For example, it assumes the same coarticulatory factor for the three formants at a certain time. This assumption does not necessarily hold.

ACKNOWLEDGEMENT

This research was partially supported by Grants 0117911 and 0082718 from the National Science Foundation. We thank Alexander Kain and John-Paul Hosom for inspiring discussions and valuable comments.

REFERENCES