# THE NEMOURS DATABASE OF DYSARTHRIC SPEECH: A PERCEPTUAL ANALYSIS

James B. Polikoff and H. Timothy Bunnell
*The Alfred I. duPont Hospital for Children, Wilmington, DE*

## ABSTRACT

This paper provides a detailed analysis of perceptual data collected from the Nemours Database of Dysarthric Speech. The database includes a collection of 740 short nonsense sentences spoken by 10 male speakers with varying degrees of dysarthria that have been marked at the phoneme level. All of the sentences were in the form "The X is Ying the Z", where X and Z were monosyllabic nouns and Ying was a monosyllabic verb root with an "ing" ending. The complete set of sentences consisted of 74 nouns and 37 verbs each produced twice by each talker. For each talker, a minimum of 5 normal hearing subjects listened to the complete set of 74 sentences 12 times for a minimum of 60 presentations of each production of each word. As subjects listened to each sentence, they chose the three words that they thought the talker was trying to produce from a set of 4 to 6 response alternatives for each word. The response alternatives were chosen such that they differed from the target word on a single phoneme. The target phonemes could occur in word initial, word final, intervocalic, or medial positions. In addition to word position, target phonemes for the perceptual analysis were classified by voicing (voiced or unvoiced), primary articulator (lips/jaw, tongue tip, tongue body, glottis), and manner of articulation (stop, fricative, affricate, approximate, nasal, and null).

## 1. INTRODUCTION

The dysarthrias are a family of neurologically based speech disorders. The intelligibility of dysarthric speech can range from near normal to unintelligible, depending on the severity of the dysarthria. Persons with moderate to severe dysarthria can be particularly difficult to understand especially for people who are unfamiliar with the speaker. Traditional speech therapy can be beneficial in improving the speech, but because of the neurological nature of the disorder, therapy alone cannot be expected to correct the speech to near "normal" quality. For these talkers, it would be desirable to have a speech prosthesis that could record their speech as input, process the speech, and then produce more intelligible speech as output. The Nemours database was compiled in order to collect baseline data on the segmental intelligibility of dysarthric speech as a first step in an effort to develop such a prosthesis. The database consists of a controlled set of high quality speech samples recorded from eleven male speakers with varying degrees of dysarthria. A detailed perceptual analysis of the segmental intelligibility from ten of the eleven speakers was conducted by having normal-hearing college students listen to the speech in a controlled experimental setting and choose the words they thought the speaker was saying from a closed response set.

## 2. METHOD

### 2.1. Dysarthric Speakers and Sentence Material

Because we decided to concentrate first on issues of segmental intelligibility, we constructed a list of words and associated foils that would allow us to capture specific phoneme production errors in a manner similar to that described by Kent, Weismer, Kent, and Rosenbek (1989). Each word in the list (e.g., boat) is associated with a number of (usually) minimally different foils (e.g., vote, moat, goat). The word and its associated foils form a closed response set from which listeners in a word identification task must select a response given a dysarthric production of the target word. However, unlike the test designed by Kent, et al. (1989) we embedded the test words in short semantically anomalous sentences, with three test words per sentence (e.g., The boat is reaping the time). Note also that, unlike Kent, et al. (1989) who used exclusively monosyllabic words, we have included bisyllabic verbs in which the final consonant of the first syllable of the verb can be the phoneme of interest. That is, the /p/ of reaping could be tested with foils such as reading and reeking. The complete stimulus set consists of 74 monosyllabic nouns and 37 bisyllabic verbs embedded in sentences (there are two nouns and one verb per sentence). To counterbalance the effect of position within the sentence for the nouns, we had talkers record 74 sentences with the first 37 sentences randomly generated from the stimulus word list, and the second 37 sentences constructed by swapping the first and second nouns in each of the first 37 sentences.

The talkers were eleven young adult males with dysarthrias resulting from either Cerebral Palsy or head trauma. Seven of the talkers had Cerebral Palsy. Of these seven, Three had spastic CP with quadriplegia, two had athetoid CP (one quadriplegic), and two had mixed spastic and athetoid CP with quadriplegia. The remaining four talkers were victims of head trauma (one quadriplegic and one with spastic quadriparesis), with cognitive function ranging between Level VI-VII on the Rancho Scale. The speech from one of the talkers (head trauma, quadriplegic) was extremely unintelligible. Because the speech was so poor, it was not marked at the phoneme level, and perceptual data were not collected for this talker.

The recording sessions were conducted in a wheelchair accessible sound-attenuated booth using a table-mounted Electrovoice RE55 dynamic omni-directional microphone connected to a Sony digital audio tape recorder, model PCM-2500 situated outside the recording booth. The talker was seated, typically in a wheelchair, next to the experimenter or

speech pathologist, and approximately 12 inches from the microphone. The recording sessions began with a brief examination by a speech pathologist including administration of the Frenchay Dysarthria Assessment (Enderby, 1983). Following the assessment and after a short break, the experimenter entered the recording booth to lead the talker through a series of recordings which included the set of 74 semantically anomalous sentences described above followed by two paragraphs of connected speech. The speech material was typed in large print on paper placed in front of the talker and the talker was given some time to familiarize himself with it before the recording began. For the sentence material, each sentence was read first by the experimenter and then repeated by the talker. This assisted all talkers in pronunciation of words and was essential for some subjects with limited eyesight or literacy. Finally, the talkers recorded two paragraphs of connected speech. On average the entire recording session was completed in two and one half to three hours, including time for breaks. The recorded sentences of both the dysarthric talker and the experimenter were later digitized and the six words in each sentence were marked using a waveform/spectrogram display and editing program (Bunnell, 1992). Additional marks at the phoneme level have also been applied using a Discrete Hidden Markov Model (DHMM) labeler (Menéndez-Pidal, et al, 1996).

## 2.2. Listeners and Identification Testing

A minimum of five normal hearing listeners were recruited from students at the University of Delaware for listening tests with each of the dysarthric speakers. Listeners were seated in a sound dampened room facing a touch screen terminal and heard sentences presented binaurally over TDH-49 headphones at an average level of 72 dB SPL.

The sentences were presented in either their original form, or in a time-warped version. The time-warped sentences were adjusted to match the timing of a corresponding normal-speech template and were typically about half the duration of the original speech. The mode of presentation was random within a set of trials with the constraint that half of the presentations were in original mode and half were time-warped. The presentation order of the pre- recorded sentences was also randomized. The data for the temporally-adjusted speech, however, will not be discussed in this paper.

At the start of each trial, the terminal screen was cleared and a new sentence frame appeared with the three target word locations in each sentence containing a list of possible response words from which listeners attempted to select the words that they thought the talker was attempting to say. For instance, a sentence might appear as follows:

|     | FIN  |    | SIPPING |     | BATH  |
|-----|------|----|---------|-----|-------|
| The | THIN | is | SINNING | the | BADGE |
|     | SIN  |    | SITTING |     | BATCH |
|     | BIN  |    | SIPPING |     | BASH  |
|     | PIN  |    |         |     | BASS  |
|     | INN  |    |         |     |       |

Thus, each target word was associated with several similar sounding foils and the listener had to pick the correct alternative from the list (depending on the target word, anywhere from four to six alternatives were available). Subjects indicated which words they thought they heard from the set of choices displayed by touching that alternative on the touch sensitive screen of a CRT in the sound booth. On each trial, the order of response alternatives for the target words was random.

There were two sets of 37 sentences from each talker. The first set contained one repetition of each word from the stimulus pool and the second set contained a second repetition of each word (the initial and final nouns were swapped in each sentence). Within a set, each sentence was presented once in its original form and once in time-warped form. Each set was presented 12 times to each of 5 listeners. Thus, there were a total of 60 presentations of each production of each word in each mode. The recorded material for talker SC, however, had ten listeners, so the amount of data was doubled. Talker JF had two extra listeners with a total of 9 extra set-pair presentations between them. Each production for talker JF, therefore, was heard 69 times in each mode.

Data was recorded using a program that kept track of the mode of presentation, the sentence, the word within the sentence, the correct alternative, the subjects response and the response reaction time in milliseconds (up to 30 seconds).

## 2.3. Perceptual Data

Only the perceptual data for the dysarthric speech in its original form will be discussed in this paper. The percentage of correctly identified target phonemes for each utterance from each talker are collapsed across listeners and presentations. Except for talkers SC and JF, each data point represents the distribution of 60 responses for a single utterance (5 listeners times 12 presentations). The number of responses represented in the distributions for talkers SC and JF are 120 and 69, respectively (see above). Target phonemes were also classified on the basis of word position (word initial, word final, intervocalic, or medial), voicing (voiced or unvoiced), primary articulator (lips/jaw, tongue tip, tongue body, glottis), and manner of articulation (stop, fricative, affricate, approximate, nasal, null).

## 3. RESULTS

There was a significant main effect for consonants in monosyllabic words ($F[24,216]=3.52$; $p < .001$) (see figure 1).
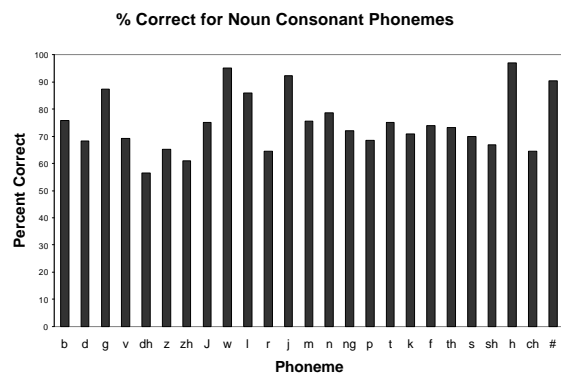


**% Correct for Noun Consonant Phonemes**

Figure 1

As expected, there was a significant main effect for vowels (F[7,63]=4.02; $p < .001$), with recognition for both front and back vowels generally higher than recognition for mid-vowels. Pairwise comparisons showed that /ɪ/ had significantly lower recognition (61.8%) than /i/ (99.0%), /o/ (89.0%), and /u/ (97.0%). As with non-disabled talkers, this may simply be because the central vowels have more neighboring vowels to be confused with than those at the extremes (see figure 2).
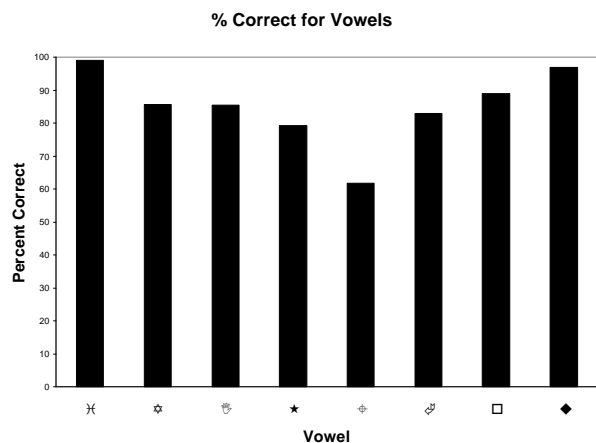
**% Correct for Vowels**



Figure 2

For monosyllabic words, word-initial phonemes were identified significantly better overall than word-final phonemes (F[1,9]=6.67; $p < .01$) (initial=76.4%, n=640; final=71.7%, n=680). This effect was more pronounced for voiced phonemes (i=80.5%, f=68.9%) than for unvoiced phonemes (i=72.9%, f=74.2%), and the effect was most apparent for phonemes produced towards the front of the mouth. Voiced phonemes produced at the lips/jaw and tongue tip were identified more accurately in the word-initial position (lips/jaw: i=81.7%, f=68.4%; tongue tip: i=79.3%, f=65.4%). Voiced phonemes produced by the tongue body were also identified better in the word-initial position, but to a lesser degree (i=80.5%, f=75.8%). For unvoiced phonemes, however, this effect was reversed. Unvoiced phonemes produced towards the front of the mouth were identified better in word-final position (lips/jaw: i=66.6%, f=80.7%; tongue tip: i=67.0%, f=76.1%), while those produced by the tongue body were identified better in the word-initial position (i=71.5%, f=63.5%).

Similar effects were found for disyllabic words. Word-initial phonemes were identified significantly better overall than intervocalic phonemes (F[1,9]=24.56; $p < .001$) (initial=78.0%, n=380; intervocalic=67.0%, n=360) and this effect was more pronounced for voiced phonemes (i=79.1%, v=62.5%) than for unvoiced phonemes (i=76.5%, v=70.5%). The interaction of voicing and position with place of articulation, however, was somewhat different from that found with the monosyllabic words. For the disyllabic words, voiced phonemes were identified better in the word-initial position than in the intervocalic position at all places of articulation, and this effect was most pronounced at the tongue tip (lips/jaw: i=89.1, v=76.4; tongue tip: i=70.0, v=40.7; tongue body: i=91.8, v=74.5; (n=0

for intervocalic other (cls/glt/null)) (see figure 3).
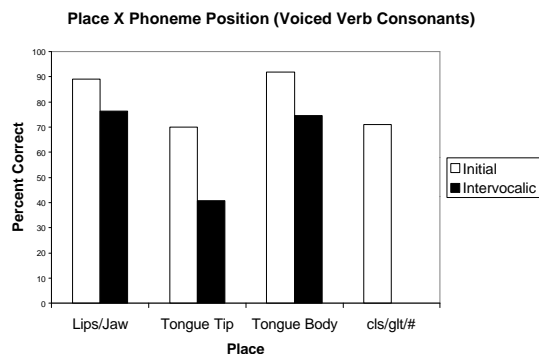
**Place X Phoneme Position (Voiced Verb Consonants)**



Figure 3

There was, in fact, a significant main effect for place of articulation (F[3,27]=20.02; p < .001), such that, tongue tip had significantly lower overall recognition (59.8%) than either lips/jaw (80.0%), tongue body (77.1%) or other (76.6%). For unvoiced phonemes, however, identification was better for phonemes produced in the intervocalic position at the lips/jaw (i=58.3%, v=78.2%) and tongue body (i=70.8%, v=78.1%), but not at the tongue tip (i=83.4%, v=44.5%) (see figure 4).

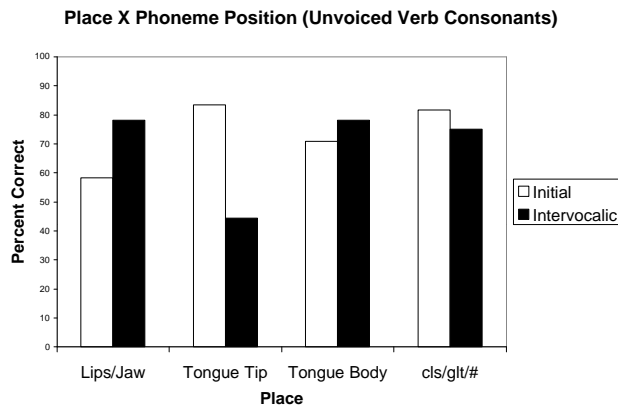**Place X Phoneme Position (Unvoiced Verb Consonants)**



Figure 4

At the tongue tip, phonemes in the intervocalic position were misidentified significantly more often than those in the word-initial position (F[1,9]=58.97; p < .001) (42.2% v. 74.4%) and this effect held true whether the phonemes were voiced or unvoiced (there were no main or interaction effects with voicing). Poor recognition for intervocalic as opposed to word-initial phonemes at the tongue tip, in fact, appears to account for the lower overall identification of phonemes produced in this location. Only the tongue-tip phonemes "d", "t", and "n" were tested in both the word-initial and intervocalic positions and each of these phonemes showed a dramatic decrease in intelligibility in the intervocalic position (d: i=65.2%, v=24.0%; t: i=86.4%, v=44.5%; n: i=93.4%, v=49.0%) (see figure 5).

The low recognition of intervocalic phonemes produced at the tongue tip suggests that the dysarthric talkers were having difficulty producing flaps and/or were having general difficulty

with coordination and coarticulation at the tongue tip.

For monosyllabic words there was a main effect of manner (F[5,45]=5.93, *p* < .001) such that null manner (90.4%, n=60)

**Tongue Tip Phonemes X Position in Disyllabic Words**

Figure 5
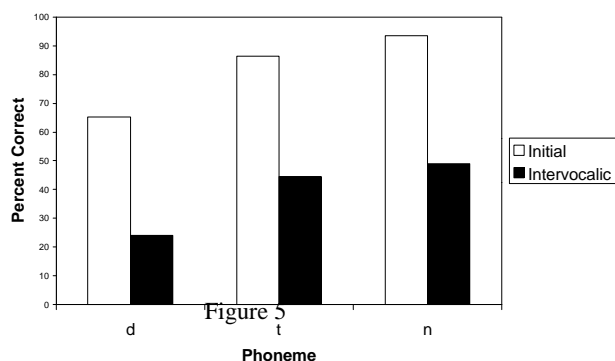
was identified significantly more often than stops (73.4%, n=440), fricatives (70.2%, n=440), affricates (69.2%, n=120) and nasals (76.4%, n=160). Approximates (84.8%, n=100) were recognized significantly better than stops, fricatives, and affricates.

Nasals, null-consonants, and fricatives in monosyllabic words were identified better in word-initial position than in word-final position (nasal: i=88.7%, f=69.0%; null: i=99.4%, f=86.0%; fric: i=73.4%, f=67.6%), while affricates and approximates were more easily recognized in the word-final position (aff: i=67.9%, f=71.9%; apr: i=83.2%, f=91.3%) Stops were identified about equally in either phoneme position (i=74.0%, f=73.0%).

Both voiced and unvoiced fricatives were identified more easily in the word-initial position in monosyllabic words (voiced: i=66.8%, f=64.3%; unvoiced: i=75.0%, u=70.0%). For stops, recognition was higher in the word-initial position for voiced stops (i=83.1%, f=64.9%) and higher in the word-final position for unvoiced stops (i=64.8%, f=77.0%). This situation was reversed for affricates (voiced: i=69.9%, f=85.2%; unvoiced: i=65.8%, f=58.7%).

For disyllabic words there was a main effect of manner (F[5,45]=6.86; p < .001) such that null manner (89.8%, n=40) was correctly identified significantly more often than stops (68.9%, n=280) and affricates (66.1, n=60).

Phonemes in all manners of articulation in disyllabic words were perceived better on average in word-initial position than in intervocalic position (null: i=98.3%, v=81.2%; affricate: i=70.0%, v=64.1%; aproximate: i=74.4%, v=(n=0); fricative: i=78.7%, v=73.2%; nasal: i=90.9%, v=62.3%; stop: i=74.8%, v=64.5%).

For those manners of articulation that had representative phonemes in both voiced and unvoiced conditions, fricatives were perceived better when voiced (v=76.4%, uv=74.9%), while stops and affricates were perceived better when unvoiced (stops: v=66.4%, uv=69.9%; affricates: v=60.1%, uv=69.1%).

## 4. CONCLUSION

The analysis showed that intervocalic and word-final consonants were correctly identified less frequently than word-initial consonants and, in agreement with previous research with similar populations, this effect was found to be more pronounced when the consonants were voiced rather than unvoiced. For the verbs, consonants produced with the tongue tip had lower overall recognition than those produced with other articulators and this effect appeared to be primarily due to sharply lower recognition rates for consonants involving the tongue-tip in intervocalic position. The low recognition of intervocalic consonants involving the tongue tip suggested that the dysarthric talkers were having difficulty producing rapid movement such as flaps and/or were having general difficulty with coordination of gestures involving the tongue tip as an articulator.

## REFERENCES
Bunnell, H.T., and Mohammed, O. (1992). "EDWave – A PC-based Program for Interactive Graphical Display, Measurement and Editing of Speech and Other Signals." Software presented at the 66th Annual Meeting of the Linguistic Society of America.

Enderby, P.M. (1983) "Frenchay Dysarthria Assessment", College Hill Press.

Kent, R.D., Weismer, G., Kent, J.F., and Rosenbek, J.C. (1989). "Toward Phonetic Intelligibility Testing in Dysarthria", *Journal of Speech and Hearing Disorders*, 54, 482-499.

Menéndez-Pidal, X., Polikoff, J.B., Peters, S.M., Leonzio, J.E., and Bunnell, H.T. (1996). "The Nemours Database of Dysarthric Speech." Proceedings of the Fourth International Conference on Spoken Language Processing, October 3-6, Philadelphia, PA, USA.