Abstract

A program has been developed which uses heuristic processing of the output of a neural network to detect syllables in connected speech. The neural network was trained using 152 sentences from the TIMIT database to perform a broad (8-class) segmental classification. This is followed by a sonority-based syllable parser which locates syllable boundaries using phonological constraints and the pattern of neural network classifications over time. On test data consisting of another 144 sentences from the TIMIT database containing 1812 syllables, this parser detected 89.0% of the syllables, including syllables bounded by sonorant consonants and syllables containing syllabic consonants. The insertion error rate was 5.1%. This program is now being applied to the identification of syllable boundaries in speech from dysarthric talkers. In preliminary analyses of 18 sentences from 9 dysarthric talkers (129 syllables), the recognition rate reached 83.8%, with an insertion error rate of 3.1%. For syllables correctly detected, the RMS error in locating the syllable nucleus boundaries was 42.7 msec.

I. PROBLEM

Can we adjust the heuristic processing component to syllabify dysarthric speech while using the ANN classifier which was trained on normal speech?

II. RATIONALE

We chose this approach because:

1) Large normal speech databases are readily available for training ANNs, but dysarthric speech databases are not.

2) Adjusting only the heuristic component, if effective, would be much more efficient than retraining the ANN classifier and adjusting the heuristic component.

3) It is unclear whether an ANN if trained on poor exemplars of the broad phonetic classes (using dysarthric speech) would necessarily outperform an ANN trained on good exemplars of the same classes (using normal speech).

III. ANN CLASSIFIER

1) Training data - 152 sentences from the TIMIT database.

2) Testing Data - an additional 144 sentences from TIMIT database.

3) Broad Phonetic Classes:

S - Silence

O - Obstruent

N - Nasal

L - Lateral

R - Retroflex

W - Approximant /w/

Y - Approximant /y/

V - Vowel

Note that in preliminary training, this 8-category model consistently outperformed a 5-category model which collapsed the classes {L, R, W, Y} into a single (liquid/glide) class. Other possible breakdowns were not tested.

4) Acoustic Feature Vectors:

- 12 Bark Cepstral parameters (one frame) computed from
- 30 msec Hamming window moved in
- 10 msec Steps

5) ANN architecture:

- 108 input nodes (9 frames by 12 parameters/frame)
- 30 hidden nodes in one layer
- 8 output nodes
- fully-connected feed-forward net

(See Figures 1 and 2)

6) Classification Training:

- approximately equal number of frames selected for each class.
- frames within 15 msec of a segment boundary excluded
- parameters scaled per sentence to -1 <= X <= +1 using a linear scaling factor
- conjugate gradient optimization technique used for weight adjustment (Barnard and Cole, 1990)

7) Classification Performance:

- Overall 80.28% correct classification of each frame in test data.
- 84.88% correct excluding frames within 15 msec of a segment boundary.
- Training data classified correctly 90.1% of frames.

Figure 1: Acoustic features for ANN training case

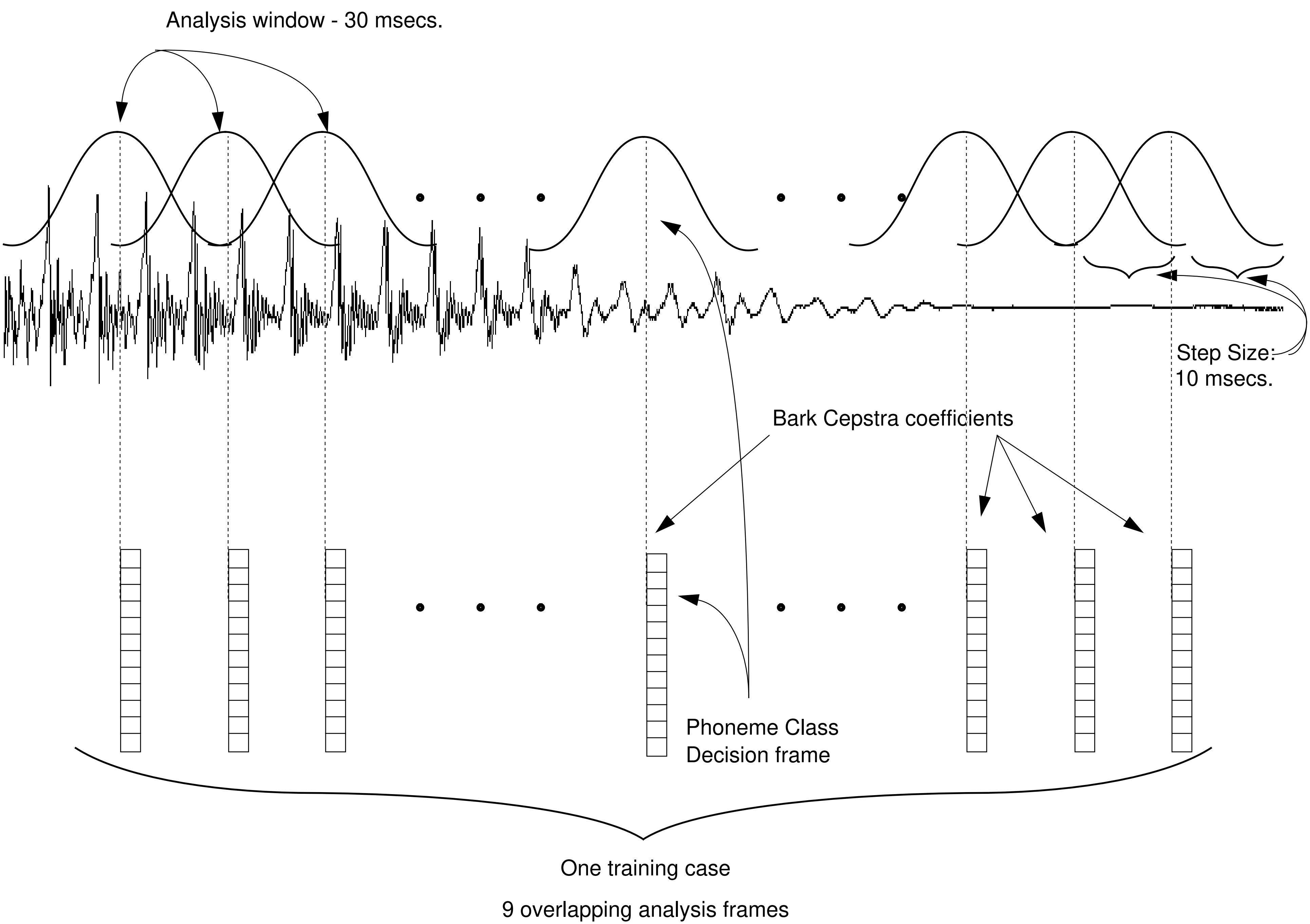
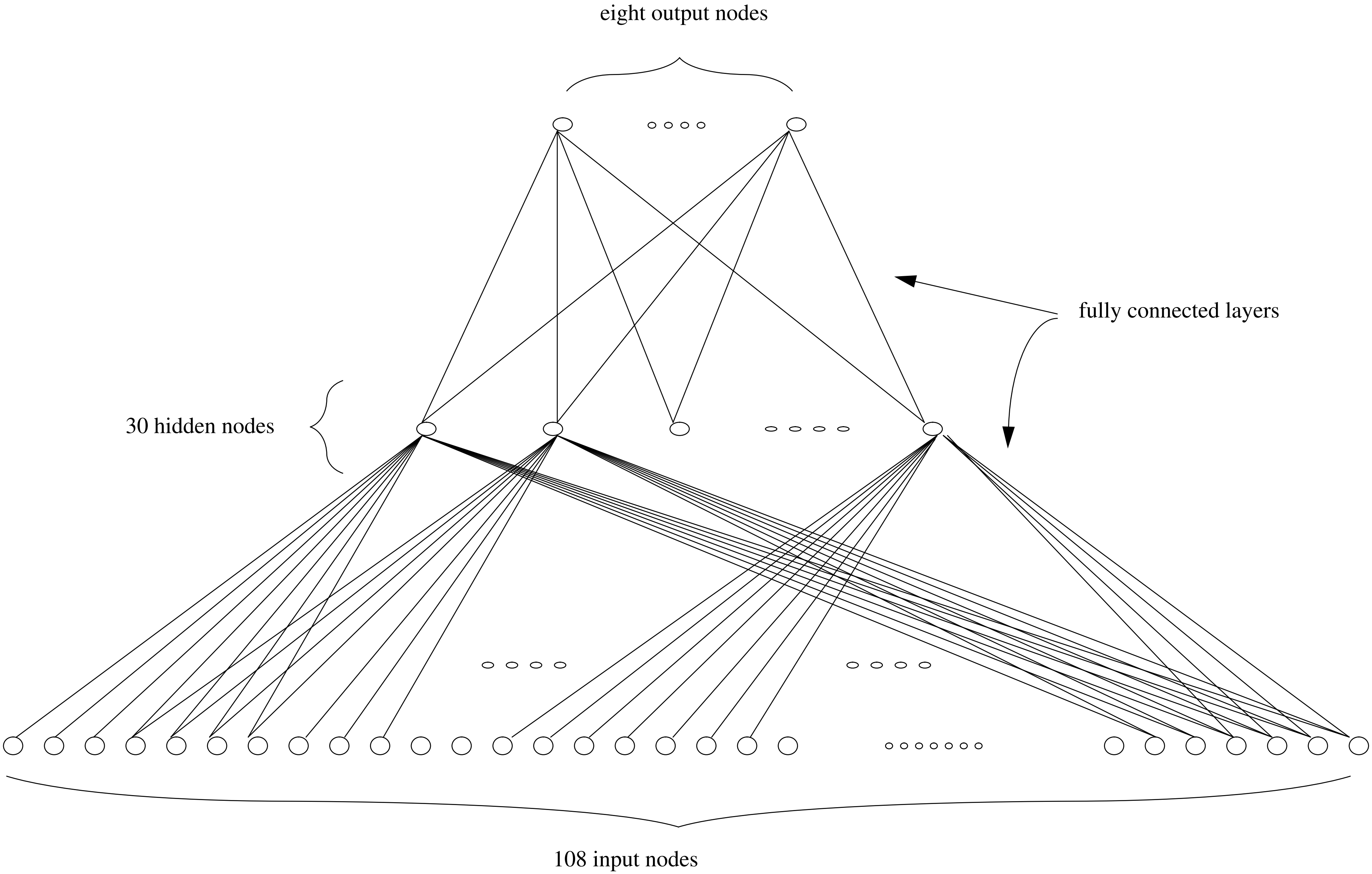


Figure 2: General layout of the network for each training case



IV. HEURISTIC SYLLABIFIER

Sonority-based syllabification proceeds in three stages from raw output of ANN.

Stage 1 - Map 8 ANN classes to 5 classes

S	-->	S
O	-->	O
N	-->	N
{L,R,W,Y}	-->	G
V	-->	V

Stage 2 - Clean output to remove improbable classifications

XYX --> XX (if duration(Y) < P)

(actual value of P used: 20)

Stage 3 - Locate syllable boundaries based on:

(i) Sonority hierarchy.

(ii) maximal onset principle.

(iii) ambisyllabicity principle.

Normal Speech Performance:

1812 syllables in the same 144 TIMIT sentences used for ANN testing.

1. Correct Detection = A syllable was detected if more than 50% of its nucleus duration was within the span of the estimated syllable.
2. Insertion - The estimated syllable could not be uniquely assigned by Criterion 1 to an actual syllable.
3. Deletion - No syllable was assigned by Criterion 1 to an actual syllable.

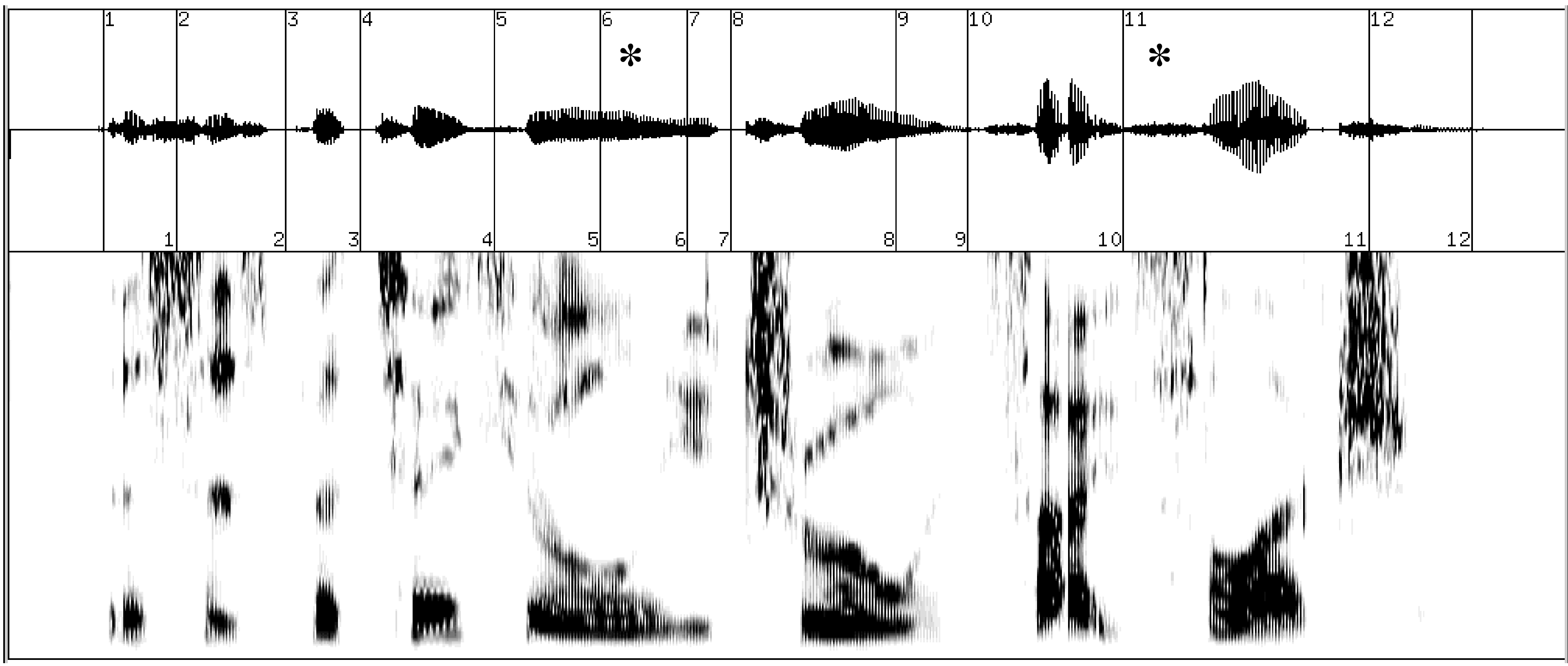
Results:

Correct Detection = 89.0%

Insertion = 5.1%

Deletion = 11.0%

Figure 3: **Example of syllabification in normal speech**
*** Region 6 - insertion error (/l/ classified as syllable nucleus)**
*** Region 11 - deletion error (syllabic /l/ merged with next vowel)**



“This is particularly true in site selection.”

TIMIT sentence dr1/fjsp0/si804/

V. ADAPTATION TO DYSARTHRIC SPEECH

Stage 1 - Map 8 ANN classes to 5 classes (unchanged)

S	-->	S
O	-->	O
N	-->	N
{L R W Y}	-->	G
V	-->	V

Stage 2 - Clean Output to remove improbable classifications

The single cleanup rule was replaced by the following set of rules applied in the order shown.

1. V-Deletion: CVC --> CC [duration(V) < Pv]
2. C-Deletion: VCV --> VV [duration(C) < Pc]
3. G-Deletion: VGV --> VV [duration(G) < Pg]

4. X-Deletion: VXV --> VV [duration(X) < Px]
(X = sequence of different Cs)

5. G-Replacement: CGC --> CVC [duration(G) > Pr]

See Figure 4 for an example.

Stage 3 - Locate syllable nucleus boundaries. Treat any G adjacent to a V as part of the V.

Figure 4: Example of rule application for removing improbable classifications

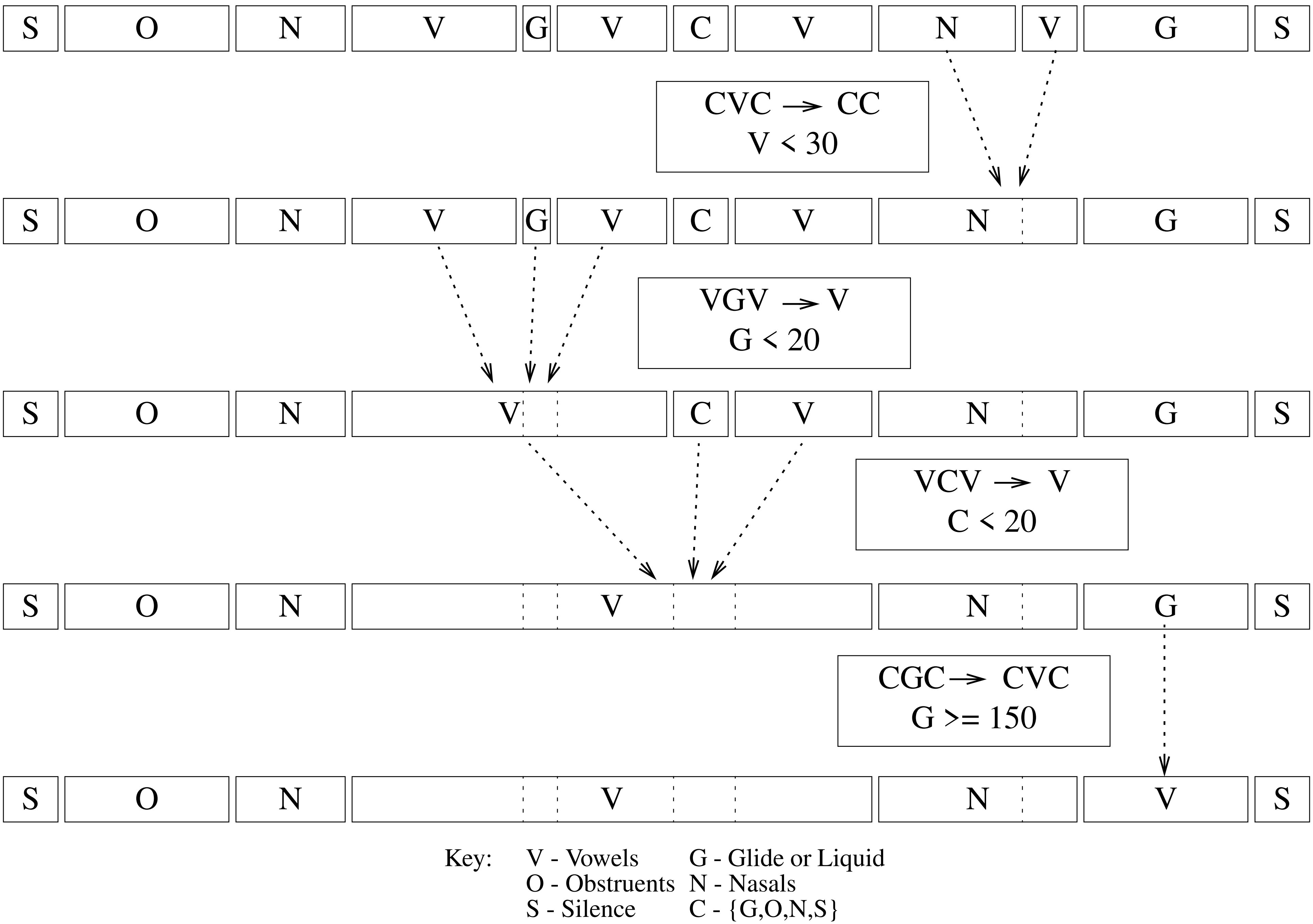
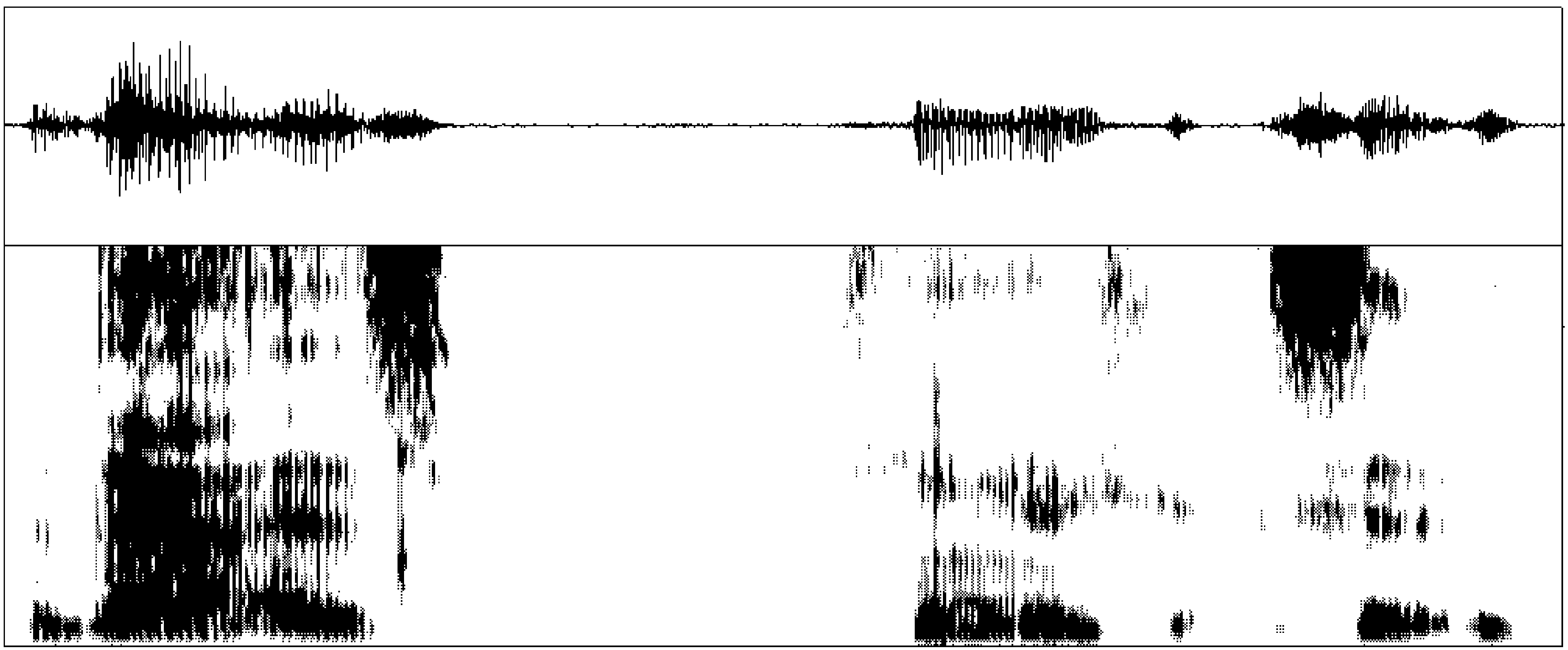
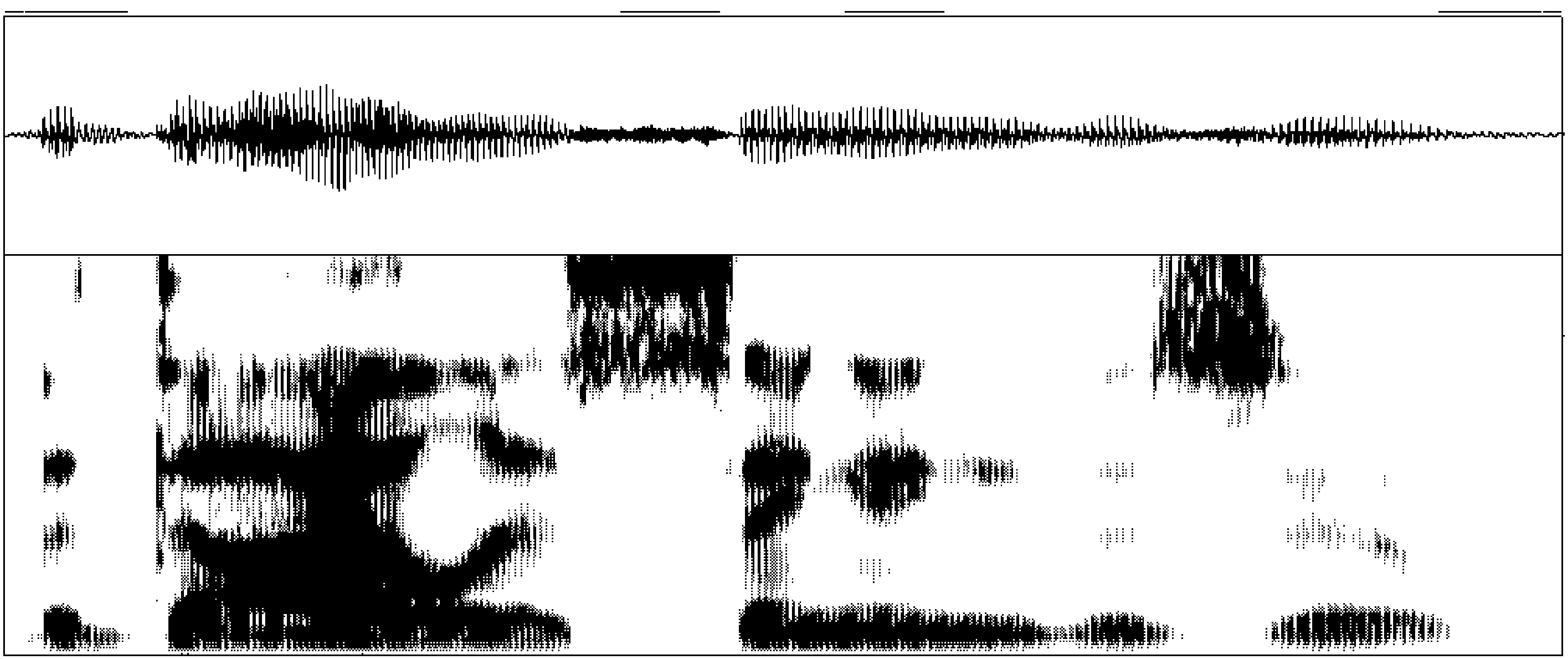


Figure 5: Example of dysarthric speech.



Dysarthric Speech - “The dial is singing the zoo.”

Figure 6: Normal exemplar of the same sentence in Figure 5.



Normal Speech - “The dial is singing the zoo.”

Tuning process:

To assign real values to each of the rule parameters (Pv - Pr), all possible combinations of the parameter values within a range believed to be useful were examined over a set of 18 sentences spoken by 9 different dysarthric talkers and comprising a total of 129 syllables.

The parameter ranges examined were:

30 <= Pv <= 70

120 <= Pr <= 200

20 <= Pg <= 70

20 <= Pc <= 70

40 <= Px <= 80

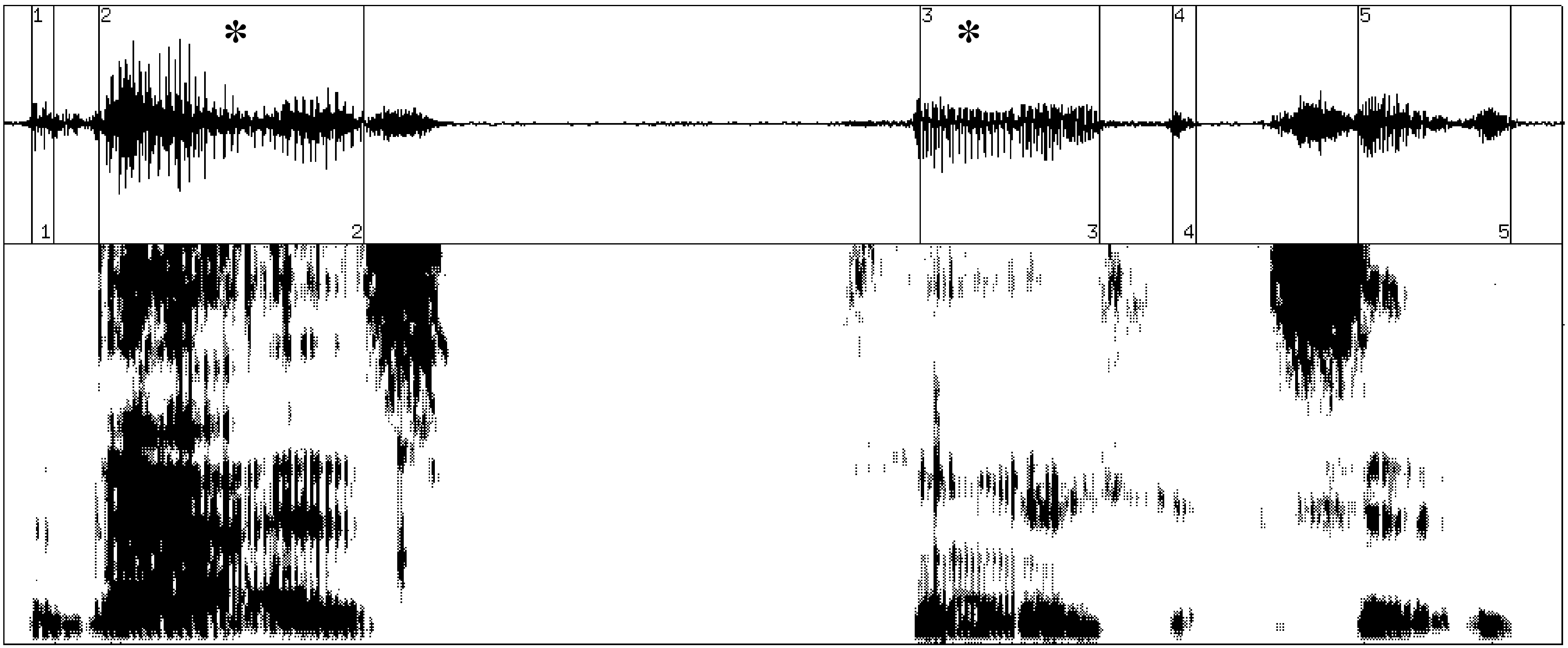
Results:

Correct Detection = 83.8%
Insertion = 3.1%
Deletion = 16.2%
RMS = 43 msec.

Parameter values used:

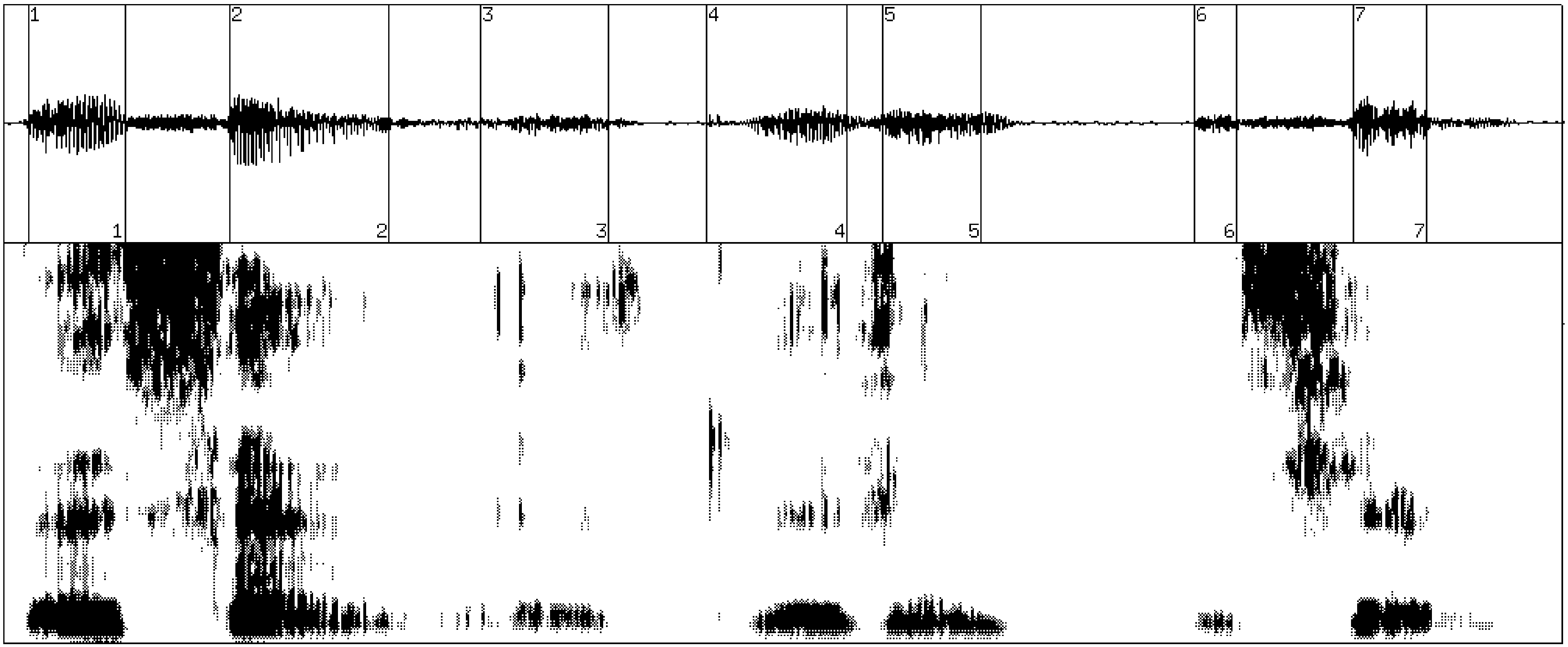
Pv = 70
Pr = 120
Pg = 40
Pc = 20
Px = 50

Figure 7: **An example of syllable nucleus detection in dysarthric speech.**
*** Regions 2 and 3 contain deletion errors (two syllables merged).**



"The dial is singing the zoo."

Figure 8: An example of syllable nucleus detection in dysarthric speech with no insertion or deletion errors.



"The sin is leaping the chew."

VI. CONCLUSIONS & FUTURE DIRECTIONS

1) This approach is encouraging - parameters chosen for characteristics of dysarthric speech led to only slightly poorer performance than for normal speech (83.8% versus 89%).

2) Evaluation of the parameters obtained for dysarthric speech using extended testing materials awaits completion of labeling in our dysarthric speech database.

3) Reasonably accurate syllabification could be part of the processing in a speech prosthesis for dysarthric talkers which would attempt to produce more normal sounding speech given dysarthric speech input.

References

Barnard, E. and Cole, R. (1990). "A Neural-Net Training Program Based on Conjugate-gradient Optimization," OGC Tech Report No. CSE 89-014.

Acknowledgments

1) Work supported by Nemours Foundation

2) The authors wish to thank Shirley Peters for her assistance in preparing this presentation.