Adjusting Dysarthric Speech Timing using Neural-Nets

Shirley M. Peters and H. Timothy Bunnell

A. I. duPont Institute Wilmington, DE

Abstract

Speech timing problems associated with dysarthria often involve the presence of periods of extraneous silence and non-speech sounds as well as inappropriately timed or misplaced speech gestures. This study evaluated the performance of neural networks in detecting the presence of inappropriate or non-speech sounds and extraneous silence. The 'opt' neural network program [E. Barnard and R. Cole. OGC Tech. Report No.CSE 89-014] which uses a conjugate gradient algorithm to adjust node weights was trained to recognize breaths and silence in a reading of the rainbow passage by a single dysarthric (Cerebral Palsy) talker. Input to the network consisted of a sequence of frames of parameters derived from spectral analysis of the speech. The output was a binary (speech/non-speech) decision for the segment of signal corresponding to the middle frame of the input sequence. Networks of various size and configuration were trained on half the available data and tested on the remaining data.

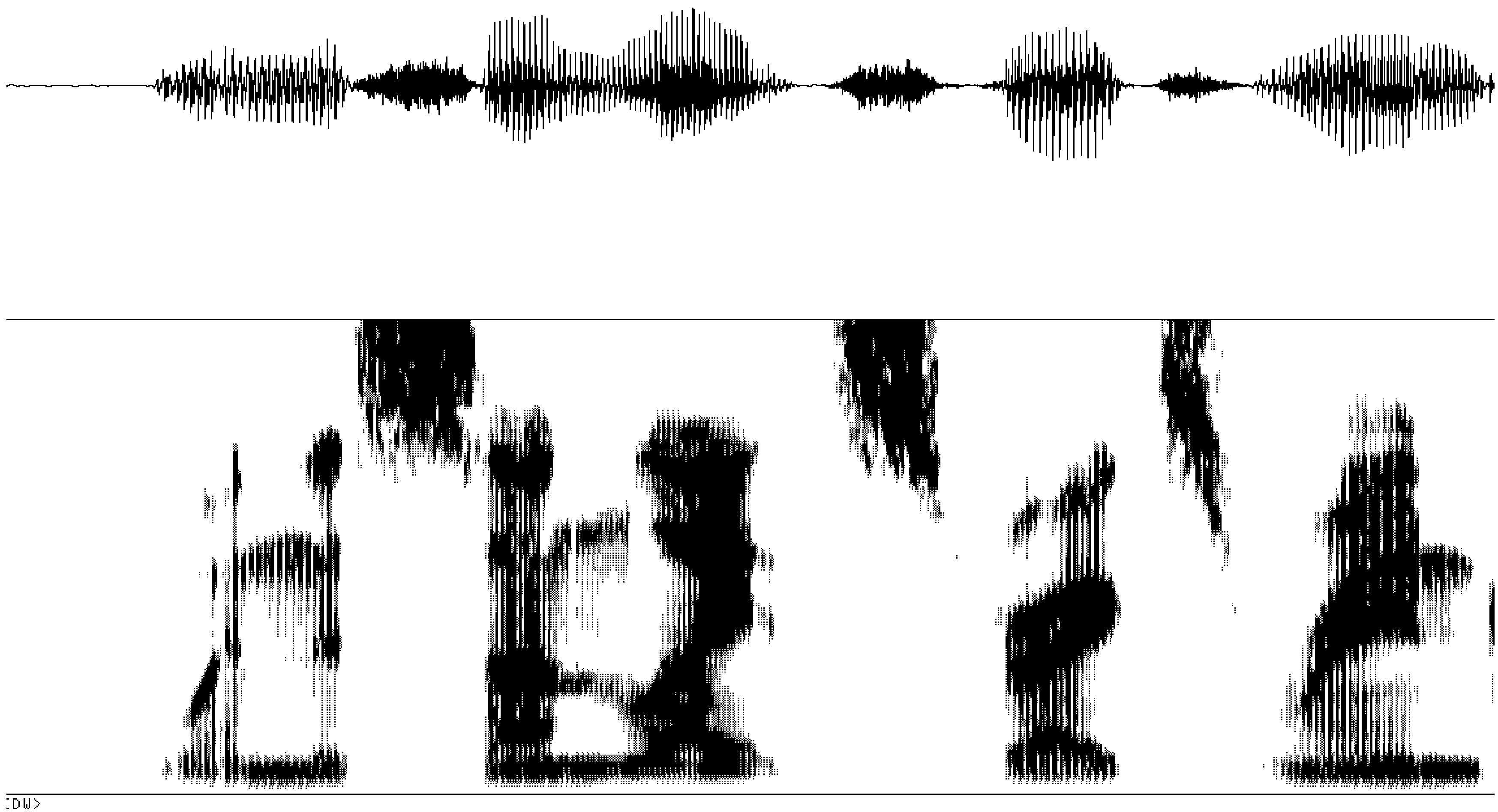
Introduction

Dysarthria is a neurological speech disorder with characteristics such as poor control of individual articulator movements and poor coordination between articulators. Even with the benefits of speech training or therapy, moderate to severe dysarthrias can significantly hinder speech communication. We are exploring the application of speech processing techniques to the development of a communication aid for dysarthric talkers which would improve the intelligibility of their speech.

One problem area to be addressed is that of disordered speech timing. Dysarthrias frequently result in periods of extraneous silence and non-speech sounds as well as inappropriately timed speech gestures. Some of these effects are illustrated in Figure 1. This Figure shows the waveform and spectrogram of the phrase “when the sunlight strikes rain” as spoken by a normal adult male talker (Figure 1.1) and an adult male dysarthric talker (Figure 1.2). The dysarthric speech contains long silences both in and between words and audible breaths which can have a speech-like formant structure. The formant structure of the breaths make it difficult to distinguish them from actual speech sounds.

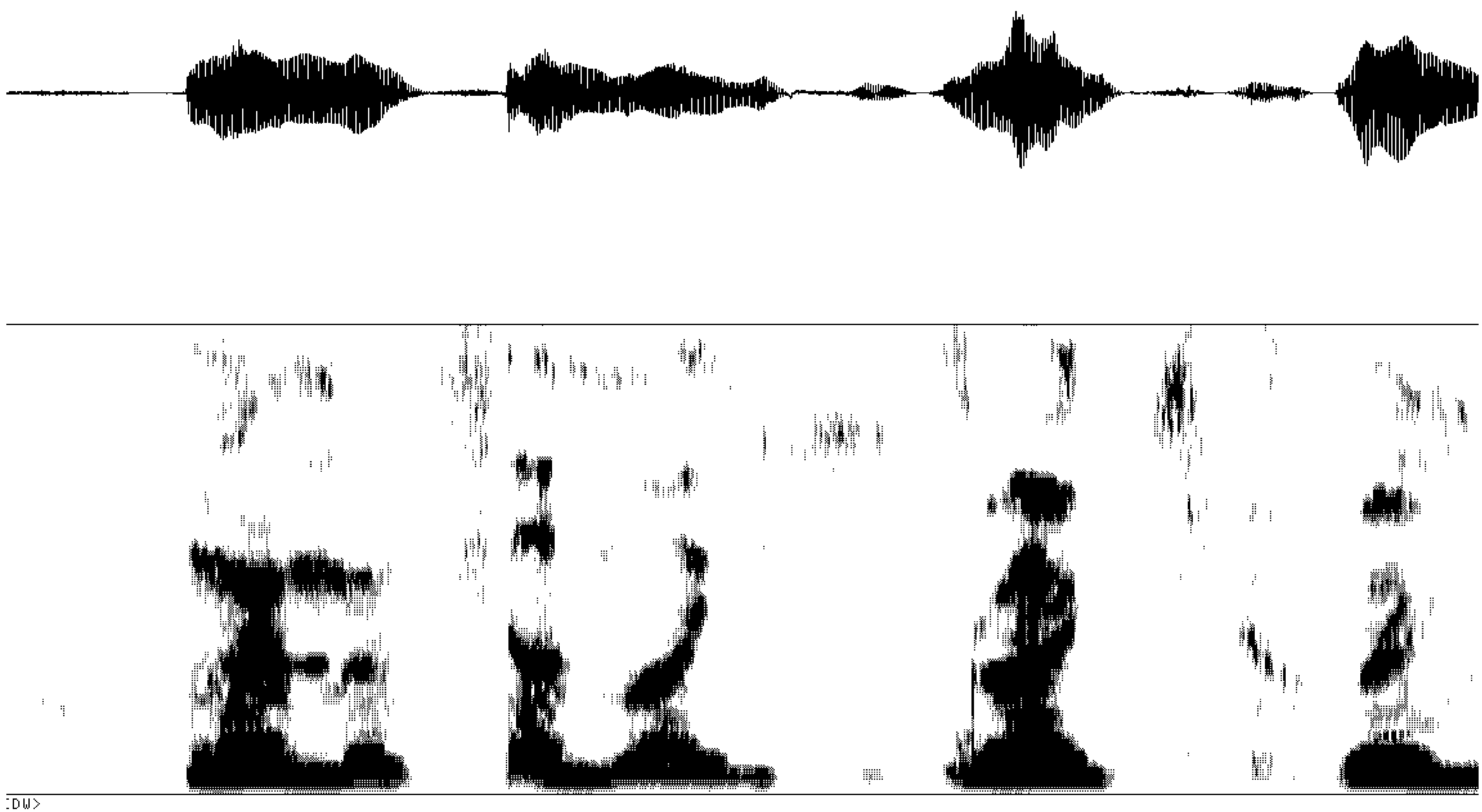
The present study evaluated the performance of several neural networks in detecting the presence of inappropriate or non-speech sounds and some classes of extraneous silence in the speech of a single adult male dysarthric talker.

Figure 1.1: The differences between dysarthric and normal speech are apparent in these waveforms and spectrograms of the speech segment “when the sunlight strikes rain”. The normal speech has very distinct differences between fricatives and voiced speech. The dysarthric speech, however, has energy appearing in the low frequencies during some of the fricatives, and audible breaths that have voice-like formant structures.



Normal Speech - “when the sunlight strikes rain” - 1781.3 msecs

Figure 1.2:



Dysarthric Speech - “when the sunlight strikes rain” - 4154.7 msecs.

Method

The following outline describes the methods used to evaluate the performance of neural networks in detecting the presence of inappropriate or non-speech sounds from an adult male dysarthric talker.

- Speech: The rainbow passage.
- Talker: One adult male with Cerebral Palsy.
- Speech Digitization: 16 kHz sampling rate with 16-bit precision
- Pre-processing: (See Figure 2.)

  "Principal Components" spectral representation.

    Speech was preprocessed by computing the first six terms of the cosine expansion of the Bark scaled log magnitude spectrum. These components are closely related to principal components derived from a factor analysis of vowel spectra (e.g., Pols, et. al, 1969), and are referred to as such in this study.

  32 Msec hamming window

    Analysis frames consisting of the six PC values were computed from 32 msec hamming windowed segments of speech.

  Step size

    The analysis frame was advanced through the speech at a rate of 5 or 10 msecs per frame.

- Neural Network

  The 'opt' neural network program which uses a conjugate gradient optimization technique to adjust the connection weights (Barnard and Cole, 1990)

- Network configuration: (See Figure 3.)

  Hidden layer

    Three different network configurations were used for training.

    Hidden nodes - 5, 10 and 20

  Output layer

    The output layer consisted of two nodes representing a binary (speech/non-speech) decision of the middle frame of the input sequence.

- Network Input: (See Figure 4.)

  Parameters varied to find optimal training set:

    Step size - 5 or 10 msecs per analysis frame

    Size of training case - 15 or 29 analysis frames

  Training and Testing cases

    A matrix of 6 PCs by 15 or 29 analysis frames of speech made up one training/testing case.

    The available speech was divided into equal halves to create the training and testing files, with enough analysis frames left out at the divisions to ensure no overlap.

  Input configurations

    Four separate input configurations were created from the step sizes and the input frame lengths.

15 analysis frames, step size 5	->	102 msec training case, 90 input nodes
29 analysis frames, step size 5	->	172 msec training case, 174 input nodes
15 analysis frames, step size 10	->	172 msec training case, 90 input nodes
29 analysis frames, step size 10	->	312 msec training case, 174 input nodes

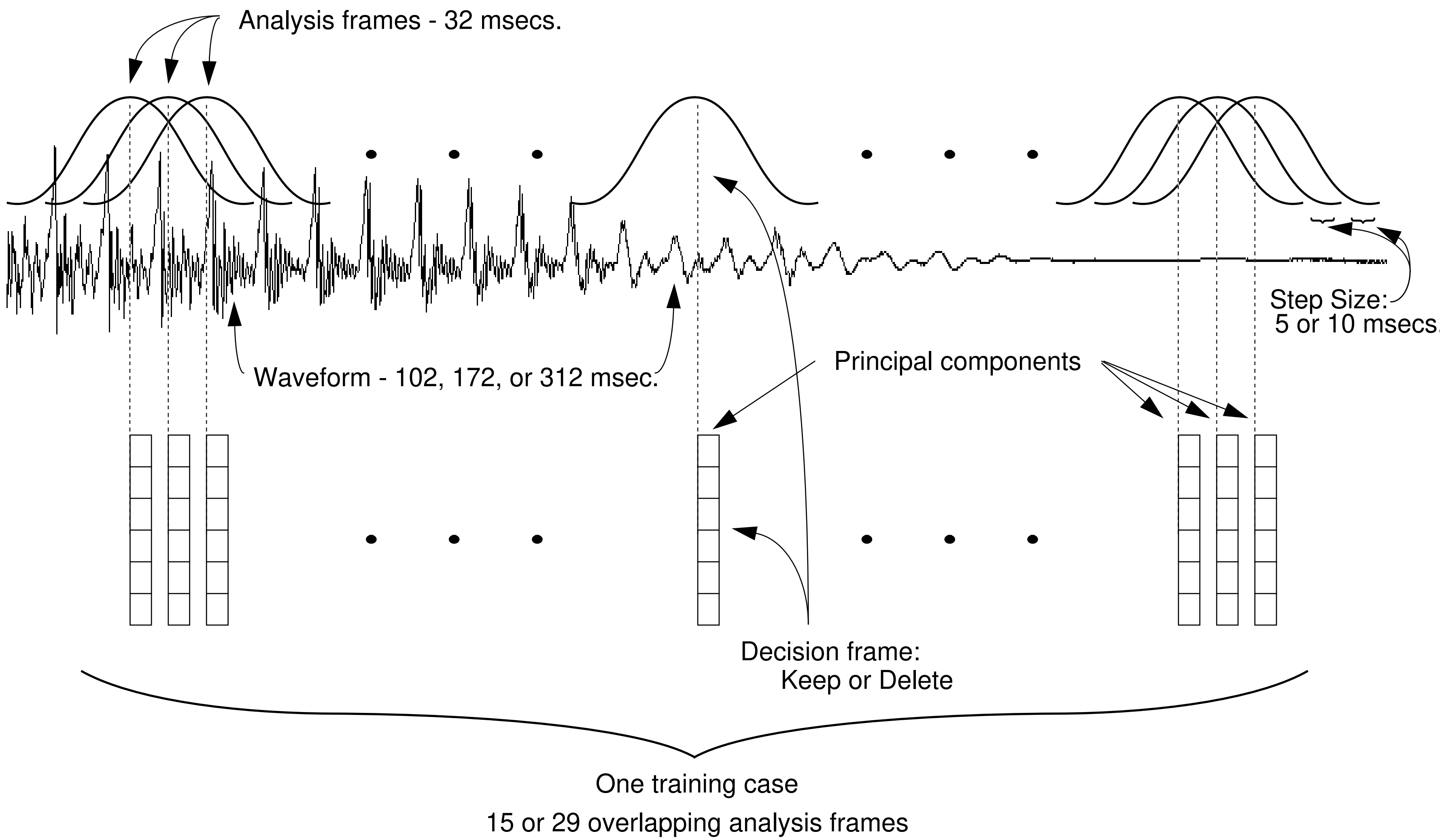
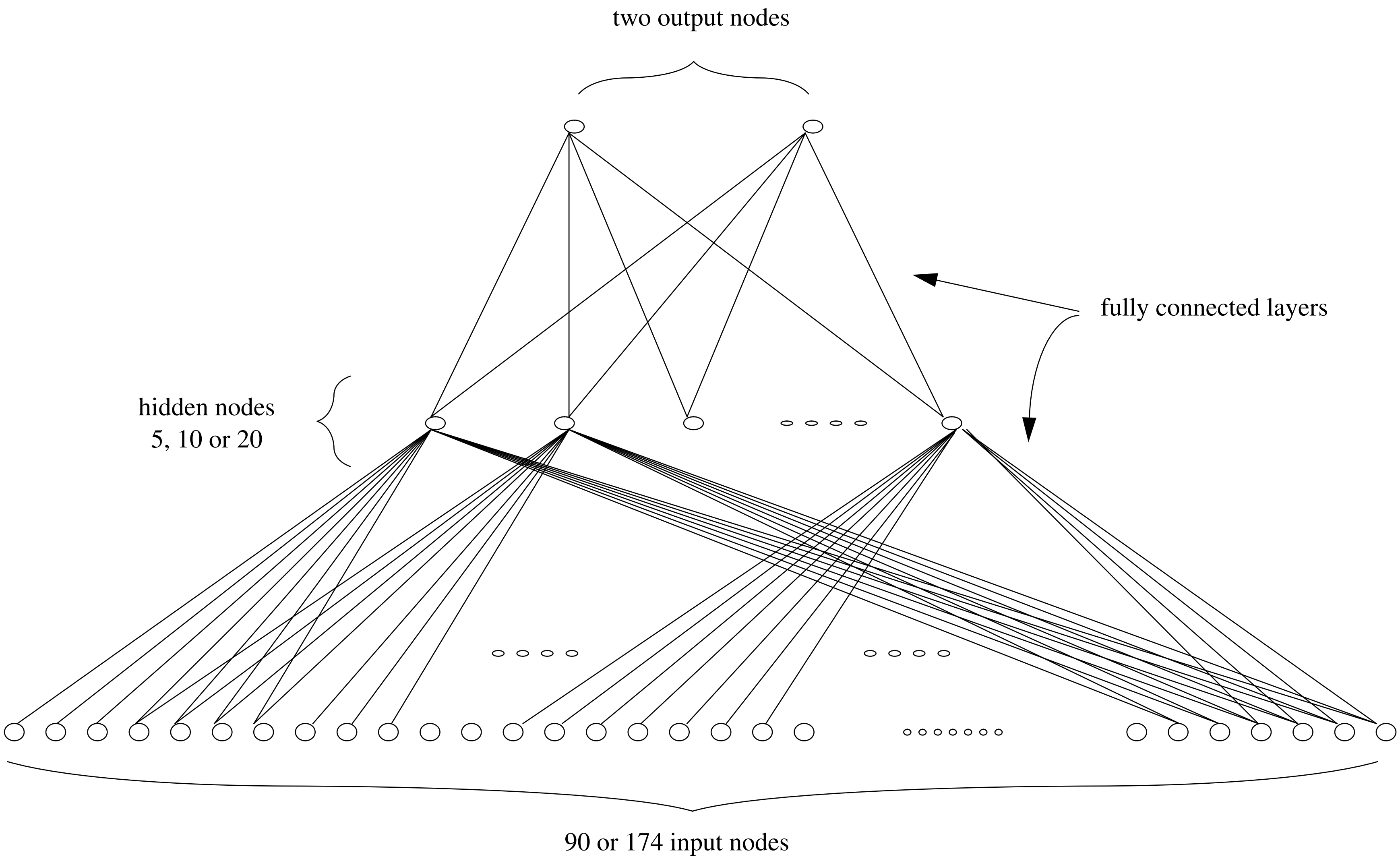


Figure 2: This example of a training case for the neural network consists of 29 overlapping spectral analysis frames, each providing 6 values describing the speech within that frame. The status (speech vs. non-speech) of the training case is determined by the status of the center frame. Each frame is the center of its own training case.

Figure 3: General layout of the network for each training case:



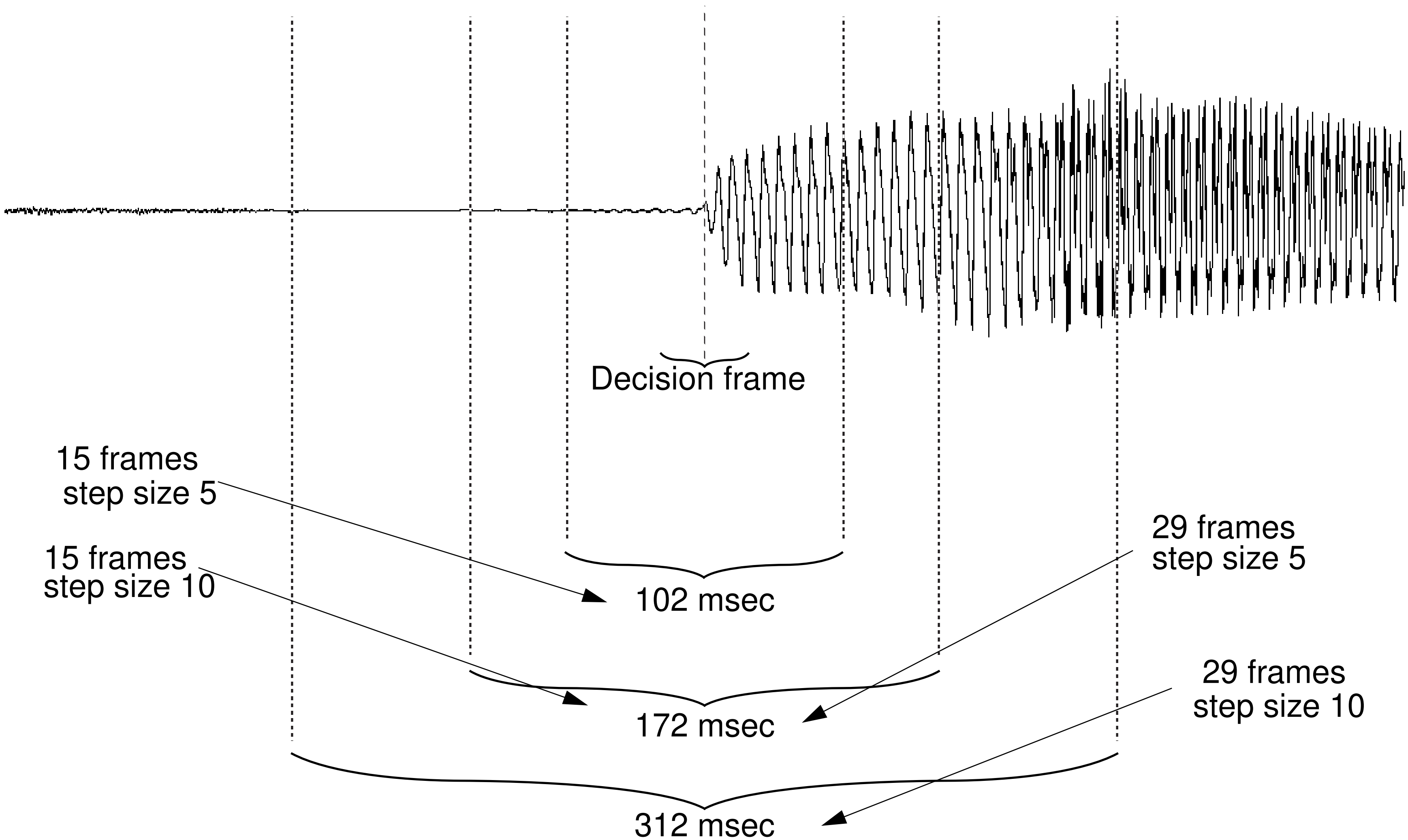


Figure 4: To determine the optimal parameters for training, the step size of the spectral analysis frames was varied between 5 and 10 milliseconds and the number of analysis frames was varied between 15 and 29 frames per training case. This resulted in four different training cases being run through the network. This waveform of the word “sunlight” shows how much surrounding information the network has to help make the decision in each case.

Table 1.1 and 1.2 show the number of cases that were available for both training and testing the networks. Half of these were used for training, the other half for testing.

Table 1.1 (5 msec step size):

		Analysis frames	
		15	29
	seconds	cases	cases
speech	52.6	10460	10432
non-speech	16.1	3141	3071
total	68.7	13601	13503

Table 1.2 (10 msec step size):

		Analysis frames	
		15	29
	seconds	cases	cases
speech	52.6	5197	5159
non-speech	16.1	1557	1497
total	68.7	6754	6656

Table 2 shows the number of connections between the input and hidden nodes for each network configuration. During the training phase, weights must be calculated for each connection.

Table 2:

		Input nodes	
		90	174
Hidden nodes	20	1800	3480
	10	900	1740
	5	450	870

Results

The best network accuracy was found using a 5 msec step size and 29 32-msec spectral analysis frames, resulting in a training case of 172 msecs. Table 3 shows the results of all network configurations with these training cases. The best network configuration used 5 hidden nodes and a total of 870 weights. With these optimum parameters, the trained network gave results of an average of 97.8% accuracy on the training set and 91.4% on the test set. However, as illustrated in Figure 5, the errors made by the networks were distributed between two separate categories: an average of 52% were clear cut errors where pieces of vowels or consonants were deleted or breaths were left undeleted, the other 48% were “judgement calls” where the decision of whether or not to keep a segment was fairly arbitrary, as in the transition between speech and non-speech. By removing the latter decisions from the error count, the reported accuracy of the results is increased by 1.6%.

Percentage correct with the optimal parameters of 174 inputs nodes, 5 hidden nodes, and 2 output nodes.

Table 3:

Step size 5, Analysis frame size 29

		Training	Testing
H			
i n	20	99.8900	90.2480
d o	10	99.7480	90.0800
d d	5	97.7540	91.4460
e e			
n s			

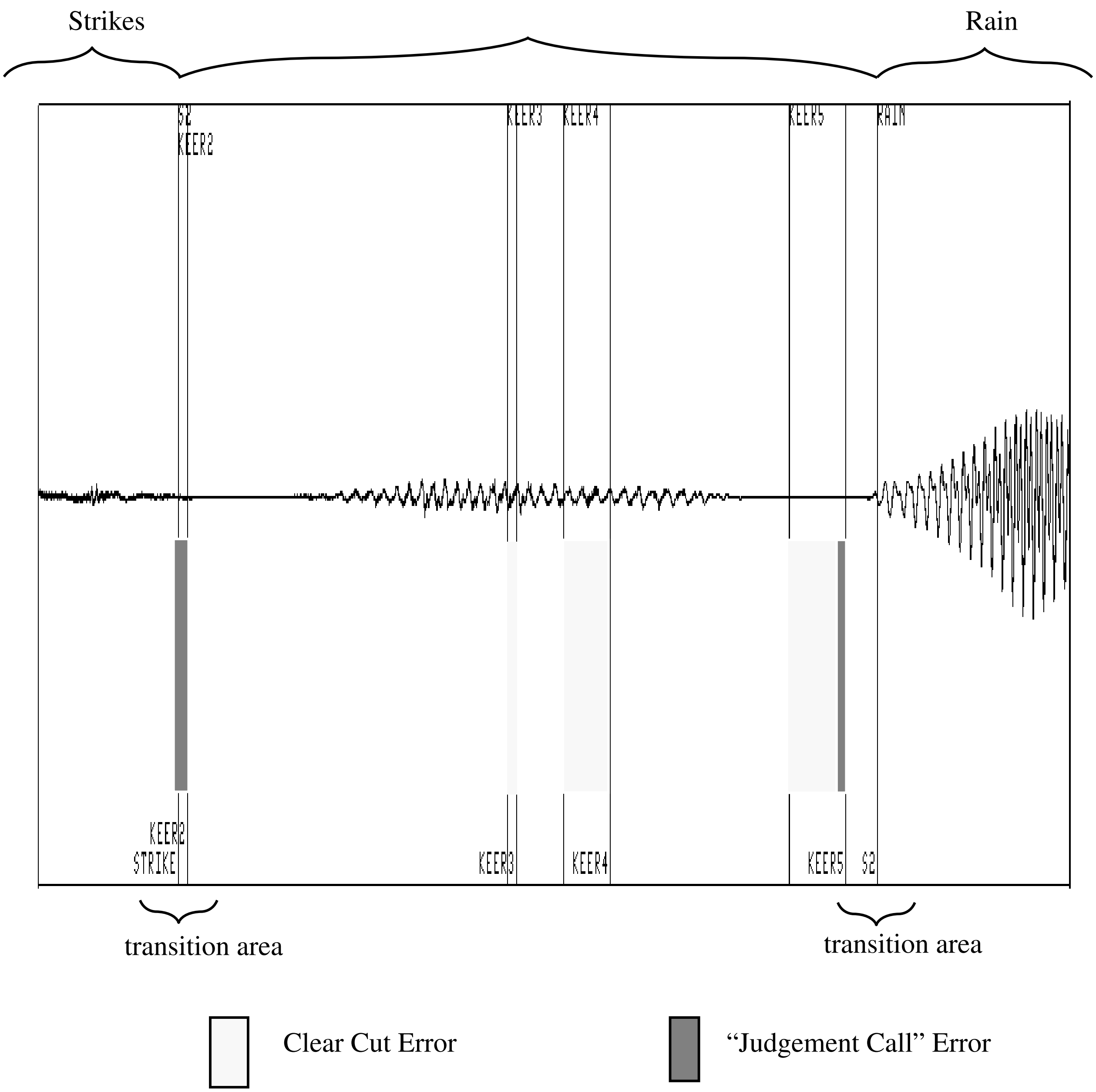


Figure 5: This waveform shows the errors made by the network. For each error in this segment, the network decided that the segment was valid speech and should be kept where the original marking showed that it should have been deleted. The light-shaded areas show clear cut errors and the dark-shaded areas are close enough to the marked transition that they could be considered judgement calls.

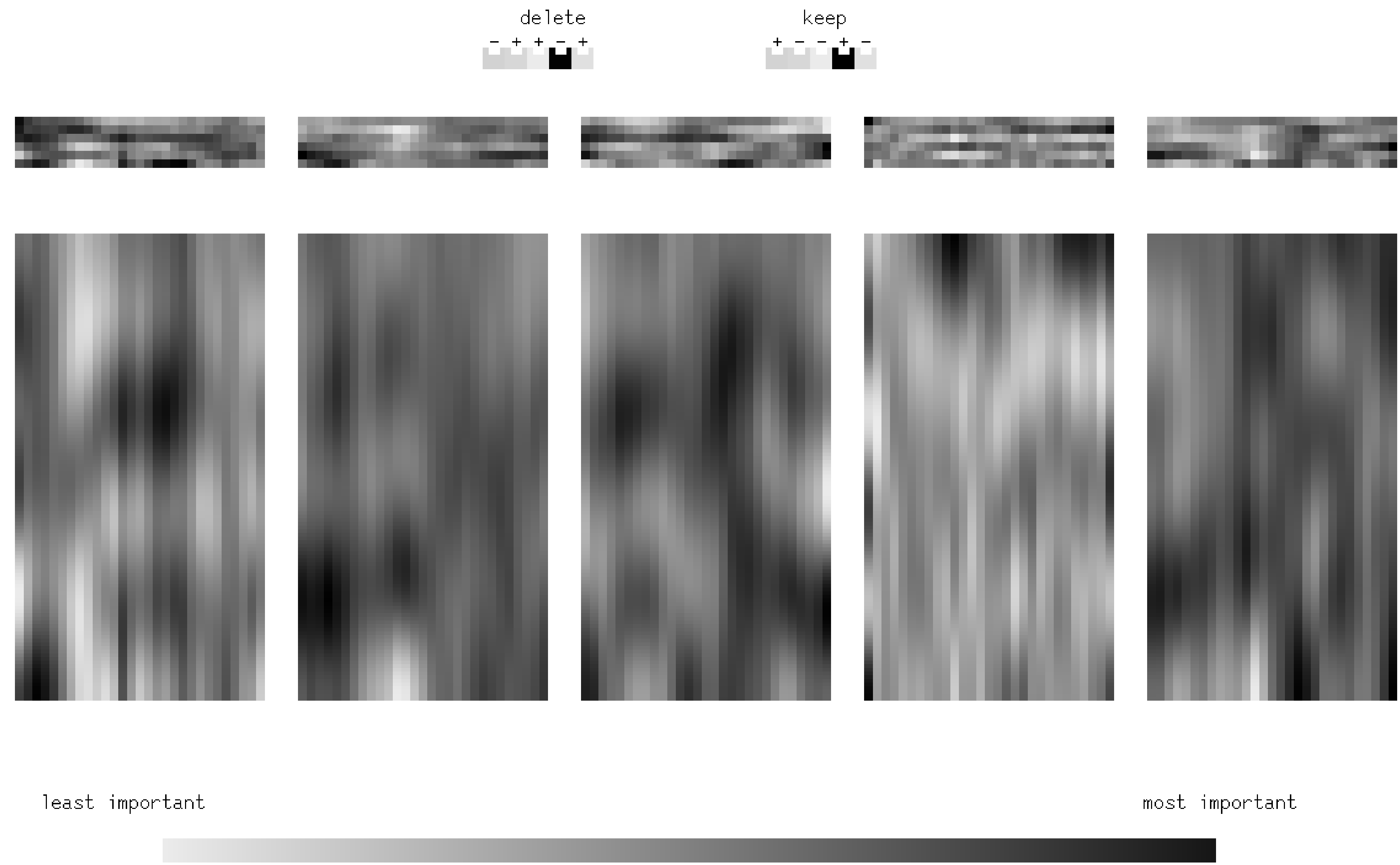
Discussion

In the present study, a neural network was trained to make the somewhat ill-defined distinction between speech and non-speech in the recording of a dysarthric talker. It is worth emphasizing that the best performing neural nets learned this distinction acceptably despite the similarity of speech and inhalation sounds (breaths), and possible inconsistencies in the training data. It is this ability of neural networks to generalize over ill-defined and even conflicting information that we feel may make them ideally suited to the classification of disordered speech characteristics. This technique could be expanded to recognize more features than just speech/non-speech. For example, any phoneme that has a duration longer than appropriate could be recognized and shortened. Specific errors of the talker could also be marked for correction. With the technology currently available, the complete recognition process could be performed in ‘real time’ and thus be used as pre-processing for further speech enhancement.

Figure 6: This is a graphical representation of the weights of the network with 172 input nodes, 5 hidden nodes and 2 output nodes. The first row shows the weights connecting the hidden layer to the output layer; the plus or minus represents how much the corresponding weight affects the decision. The darker the square is, the more important the information in the hidden node.

The weights connecting the input layer to the hidden layer are represented in the second row.

The third row shows how the weights from all the input nodes to a single hidden node would appear as a spectrogram. These weights are regarded as principal components and thus represented as spectrograms. This shows the features the network is looking for in its decision. The darker areas are those that positively affect the decision, while the lighter areas negatively affect the decision.



References:

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

Pols, L. C. W., Van Der Kamp, L. J. Th., and Plomp, R. (1969). “Perceptual and physical space of vowel sounds,” J. Acoust. Soc. Am., 46, 458-467.