UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

(1)

REPORT DOCUMENTATION PAGE

AD-A205 496

1b. RESTRICTIVE MARKINGS

DTIC FILE COPY

3. DISTRIBUTION/AVAILABILITY OF REPORT

Approved for public release; distribution is unlimited.

DOWNGRADING SCHEDULE

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION	6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
Naval Ocean Systems Center	NOSC	Naval Ocean Systems Center

6c. ADDRESS (City, State and ZIP Code)	7b. ADDRESS (City, State and ZIP Code)
San Diego, California 92152-5000	San Diego, California 92152-5000

8a. NAME OF FUNDING/SPONSORING ORGANIZATION	8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
Naval Ocean Systems Center	NOSC	

8c. ADDRESS (City, State and ZIP Code)	10. SOURCE OF FUNDING NUMBERS			
	PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	AGENCY ACCESSION NO.
San Diego, California 92152-5000	In-house			

11. TITLE (include Security Classification)

EXPLORING THE BACK-PROPAGATION NETWORK FOR SPEECH APPLICATIONS

12. PERSONAL AUTHOR(S)

S.A. Luse, D. Martin, S. Nunn, and J. Waters

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Professional Paper	FROM June 1988 TO June 1988	November1988	

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	neutral networks filtering signals speech compression

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Neural networks have sophisticated abilities for processing and filtering signals. In particular, Elman and Zipser demonstrated that the back-propagation network develops significant feature representations which may be useful for both segmenting and recognizing speech. Such networks might find applications in speech compression and/or speech normalization. The network's apparent potential for speech applications justifies further exploration, and this paper describes our work in progress.

Presented at the Navy IR/IED Symposium, 20-23 June 1988, Johns Hopkins University/Applied Physics Lab.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	UNCLASSIFIED

22a. NAME OF RESPONSIBLE PERSON	22b. TELEPHONE (Include Area Code)	22c. OFFICE SYMBOL
S.A. Luse	619-553-3652	Code 441

DD FORM 1473, 84 JAN

83 APR EDITION MAY BE USED UNTIL EXHAUSTED
ALL OTHER EDITIONS ARE OBSOLETE

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

EXPLORING THE BACK-PROPAGATION NETWORK FOR SPEECH APPLICATIONS

S. Luse
D. Martin
S. Nunn
J. Waters

Human Factors and Speech Technology Branch
Code 441
Naval Ocean Systems Center
San Diego, California

ABSTRACT

Neural networks have sophisticated abilities for processing and filtering signals. In particular, Elman and Zipser[1] demonstrated that the back-propagation network develops significant feature representations which may be useful for both segmenting and recognizing speech. Such networks might find applications in speech compression and/or speech normalization. The network's apparent potential for speech applications justifies further exploration, and this paper describes our work in progress.

BACKGROUND

Fundamental problems remain unsolved in speech processing. Although recent developments have occurred in digital signal processing techniques and their hardware implementations, many questions are unanswered concerning correct perceptual units of speech, the use of prosodic information in the speech signal, and viable voice compression algorithms.

Neural networks provide a powerful means for pursuing answers to these questions. While performing complex mappings, these networks extract fundamental information about the input signal. Relying on this ability, networks may be trained to aid in the accomplishment of difficult speech processing tasks.

One particular network paradigm, the back-propagation network, has the demonstrated capability to represent significant features of speech.[1,3] The back-propagation network is a three-layer neural network that behaves as an interpolative-associative memory. It has the ability to learn mappings by example; that is, it will generalize input-output pair relationships. The mappings that are learned are dependent on the input-output pairs that are used during the network's training period.[2]

The network's capability may be useful for isolating and modifying a variety of speech features, and might find immediate application in speech compression (to improve low data rate voice transmission), speech normalization (to improve voice recognition), and speech adaptation (for voice feature modification). In short, the back-propagation network is a powerful tool for analyzing and manipulating speech.

PURPOSE

The goal of our research is to explore how back-propagation networks, trained to learn the significant representations of preprocessed speech, affect novel speech data. Using networks of different sizes with different preprocessing methods, we hope to discover the features learned and how this information may aid the performance of difficult speech processing tasks, such as:

1. Identifying phonemes in the context of continuous speech for both speaker dependent and independent recognition;
2. Capturing speech characteristics of a particular speaker for use in speaker recognition and verification;
3. Compressing and encoding the speech signal for low data rate transmission as well as secure communications;
4. Capturing critical vocal tract parameters and speaker prosodic information such as pitch, stress, rate, etc., to be used for disguising a person's speech or changing the identity of the speaker.

We also hope to gain insight into the amount of processing power required for useful applications.

NETWORK DESIGN

The size of the back-propagation network must be designed so the network can learn to represent the fundamental aspects of the preprocessed speech. Of the three layers in our network, the input and output layers are the same size (they have the same number of processing elements), so we can perform a one-to-one mapping of speech data.

This work was performed in-house under NOSC Project No. ZT8.

The hidden layer is less than one-half the size of the input or output layer. This standard "hour glass" design forces the network to learn significant representations of speech. The network must determine these significant features to compress the data in the hidden layer and recreate it with reasonable accuracy.

Other researchers[1] have illustrated that a network, with a hidden layer size of less than one-half the size of the input layer, is capable of learning significant representations of speech. Our research will explore the use of networks with hidden layers of this size and smaller.

PREPROCESSING METHODS

A network may have difficulty drawing meaningful features from the variable speech signal, since there are practical limitations to the complexity of the mappings that a neural network can accomplish. Much research in speech recogition and perception has involved the attempt to find a representation for the speech that would simplify the problem of the great variability in the speech signal.[4] Since each representation emphasizes a different aspect of the speech signal, each will affect in a different manner a network's performance, including training time, error rate, and feature development.

Existing theories of speech recognition suggest at least three useful representations of the speech signal.[5] These representations are described below along with the preprocessing methods which can partially capture features important to the representations.

SPEECH PERCEPTION THEORY—PREPROCESSING: RAW SPEECH

Recognition can be achieved by extracting speech features, such as voice onset times and formant transitions, that have been experimentally established as being important to the human perception of speech. These time domain features are present in raw digitized speech.

The advantages of using raw speech for training the network are (1) it is easy to acquire, (2) it imposes no preprocessing perturbation of the signal, and (3) raw speech must contain all vocal tract parameters and frequency domain information in some form. No information has been eliminated. One possible disadvantage is that, since raw speech does no pre-encoding, it may place excessive burdens upon the network during training.

SPEECH RECEPTION THEORY—PREPROCESSING: FFT

The human auditory process can also be modeled by extracting parameters and classifying patterns as done in the ear, auditory nerves, and sensory feature detectors. These parameters and patterns are found principally in the frequency content of the signal.

Frequency-domain representation of speech information is doubly advantageous. First, acoustic analysis of the vocal mechanism shows the production of critical formant frequencies that permit a concise description of speech sounds.[6] Second, a great deal of evidence indicates that the ear makes a crude frequency analysis at an early stage in its processing.[6]

Furthermore, FFT data are relatively easy to extract from the time-domain signal, and there is a great deal of information in the FFT signal. However, there also is a great deal of irrelevant information in an FFT.

SPEECH PRODUCTION THEORY—PREPROCESSING: LPC

Speech can be recognized by understanding the method of speech production—the parameters describing the vocal tract. Important parameters include vocal tract resonances, rate of vibration of the vocal cords, and manner and place of articulation.

Linear predictive coding (LPC) analysis is a powerful technique for estimating the basic speech parameters, such as pitch, formants, and vocal tract area functions. LPC is based on the idea that a speech sample can be approximated with a linear combination of past speech samples. By minimizing the differences between the actual speech samples and the linearly predicted ones, a unique set of predictor coefficients can be determined.

The advantage of using LPC is that it provides an extremely accurate estimate of the speech parameters. The main disadvantage is that the data in each segment of speech have been enormously reduced to representation that contains only a few filter coefficients.

APPROACH

For our experiments, we train the back-propagation network to perform a one-to-one mapping, using speech from one male speaker. Once the network is trained to an acceptable error level, we process speech from three other male speakers as well as from the training speaker.

To determine the effect of the network, we use a speech processing system whose performance can be measured with or without the presence of the neural network (see Fig. 1). First, the system error is measured *without* the network. Digitized speech is preprocessed. Inverse preprocessing is performed and error measurements are made by comparing the input and output speech. Second, the

system error is measured *with* the network. A back-propagation network is trained to perform a one-to-one mapping of the preprocessed speech. The network output is inverse preprocessed to obtain an approximation of the input speech. Error measurements again are made by comparing the input and output speech. Together, these measurements allow us to determine the network's contribution to the system error.

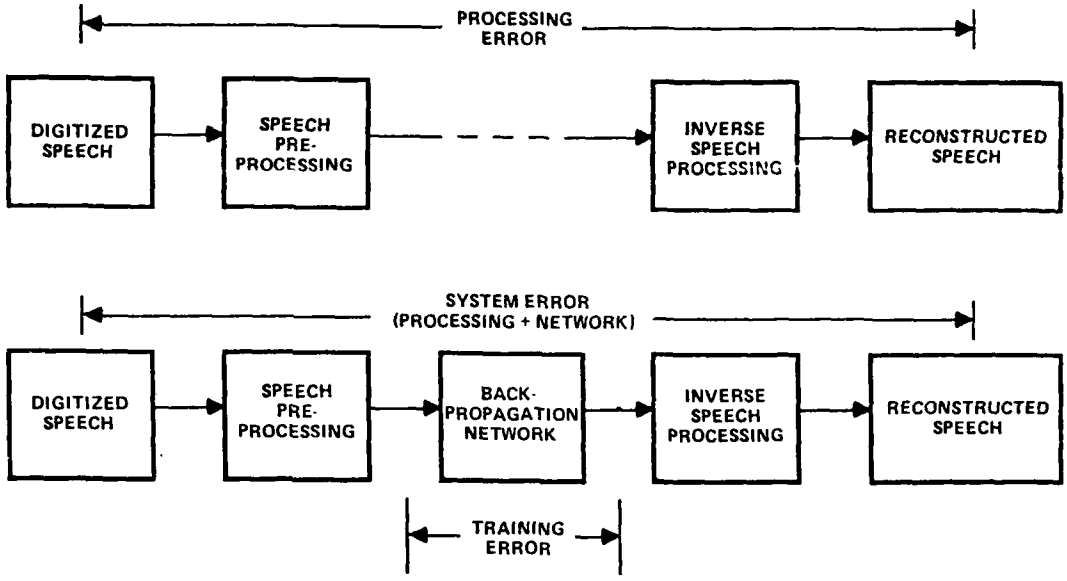

Fig. 1. System Used for Error Measurements.

With a one-to-one mapping, a mean-square error measurement can be used to compare input and output speech waveforms. Suppose the input waveform is x_n and the output waveform is y_n, both of length N samples. The error signal e_n, is defined as:

$$e_n = y_n - x_n .$$

The energy in the Error signal E^e, is defined as:

$$E_e = \frac{1}{N} \sum_n (\bar{e} - e_n)^2$$

where $\bar{e}$ is the average error over the N samples. It is useful to express the error energy as a percentage relative to the energy in the input signal S:

$$E = \frac{E_e}{S} = \frac{\frac{1}{N} \sum_n (\bar{e} - e_n)^2}{\frac{1}{N} \sum_n (\bar{x} - x_n)^2}$$

where $\bar{x}$ is the average input signal. While this error measure is not as useful as a correlation function, it is easy to calculate and provides necessary insight into the network's behavior.

For further study, we may listen to the effect of the network on novel speech and, based on the results, train new networks to emphasize such effects. We hope to design several demonstrations of potential applications to illustrate the network's capabilities.

INITIAL STUDIES

We began our studies by training a network to map raw speech, using the short word, "zero." We wanted some initial measure of training time and intelligibility. After training a 50-20-50 network (50 processing elements in the input layer, 20 processing elements in the hidden layer, and 50 processing elements in the output layer), we found training time was a matter of hours and that the processed speech was readily intelligible.

As a result of initial studies, we decided that a network size of 64-20-64 would be an appropriate starting point for our experiments. We also learned that the length of utterances had a severe impact on the required training time. When longer training inputs were tested, we determined that nonbatch processing was more efficient for our purposes (training in hours as opposed to batch training in days). We therefore chose nonbatch processing with the utterance "zero" as our test word and recorded the utterance from four male speakers with an 8-kHz sample rate. The length of the digitized waveforms varied from 4,864 samples to 5,888 samples, representing 0.61 seconds to 0.74 seconds in duration.

For raw speech, the input was segmented into 64-point blocks, each representing 16 ms of the original signal. Each block of 64-point samples was presented to the input layer.

For FFT data, the input was segmented into 128-point blocks, each representing 32 ms of the speech. After performing the FFT, each resulting 64-point block, representing the magnitude of the spectrum, was presented to the input layer. In another experiment, both FFT power spectra and phase was presented to the network.

CURRENT RESULTS

Using one speaker's utterance as a training set, we trained five back-propagation networks to within 10 percent training error. Table II shows the resulting training errors for each network.

After training the networks, we processed the training speaker data and the data for the other three speakers. The resulting errors are shown in Table III for all five networks along with the processing error for comparison.

Table II. Training Error.

	Training Errors			
Network Size	64-20-64* (%)	64-10-64* (%)	130-20-130 %)	130-10-130** (%)
Raw Speech	9.8	9.4		
FFT Speech (Power only)	9.9	10.0		
FFT Speech (Power and Phase)			9.8	—
LPC Speech***				

*Size shown is for Raw Speech. FFT Speech used networks of size 65-10-65 and 65-20-65.
**Network would not train in reasonable time and was, therefore, not included in this study.
***(Work-in-progress)

Table III. Comparing Output with Input Speech.

		Error Between *Input* and Output Speech				
		Processing Error	System Error (Processing + Network)			
Network Size		(%)	64-20-64* (%)	64-10-64* (%)	130-20-130 (%)	130-10-130** (%)
Raw	Training Speaker	0.002	9.8	9.5		
	Speaker 2	0.002	28.1	30.5		
	Speaker 3	0.003	31.0	33.6		
	Speaker 4	0.002	26.2	29.2		
FFT (Power Only)	Training Speaker	0.002	23.3	22.6		
	Speaker 2	0.002	40.5	42.3		
	Speaker 3	0.003	37.1	39.7		
	Speaker 4	0.002	46.0	47.6		
FFT (Power and Phase)	Training Speaker	0.002			52.8	—
	Speaker 2	0.002			132.0	—
	Speaker 3	0.003			169.0	—
	Speaker 4	0.002			119.0	—
LPC Speech***						

*Size shown is for Raw Speech. FFT Speech used networks of size 65-10-65 and 65-20-65.
**Network would not train in reasonable time and was, therefore, not included in this study.
***(Work-in-progress)

After this initial test, we experimented with network training using batch processing—the network interconnection weights were adjusted only after all the input patterns in the word "zero" were processed, rather than adjusting the weights after each pattern. Batch processing provided a more reliable training error measurement, although training was slower.

Using batch processing, we trained networks of various sizes on the word "zero." Rough training error measurements are shown in Table I.

Table I. Initial Studies.

Training Error vs. Time		
Network Size	Error 10 hrs (%)	Error 20 hrs (%)
64-20-64	14	10.5*
128-20-128	14	11.5
40-30-40	17	9.7*
64-40-64	13	8.0
128-40-128	8	3.5
128-60-128	8	4.23
192-60-192	7	2.3*
*Estimated		

In general, the larger the network, the lower the error. It's difficult to say whether the larger networks performed better because they received more significantly-sized chunks of speech data, or simply because they were larger and had more connections in which to store more information.

Our initial training studies verified the conclusion reached by other researchers that the back-propagation learning rate follows a power-law relation. The first graph in Fig. 2 shows pass error vs. time for seven networks of different sizes, while the second graph shows regression lines for the same data. The correlation coefficient for all regression lines is 0.96 or greater, indicating a strong fit to the power-law.

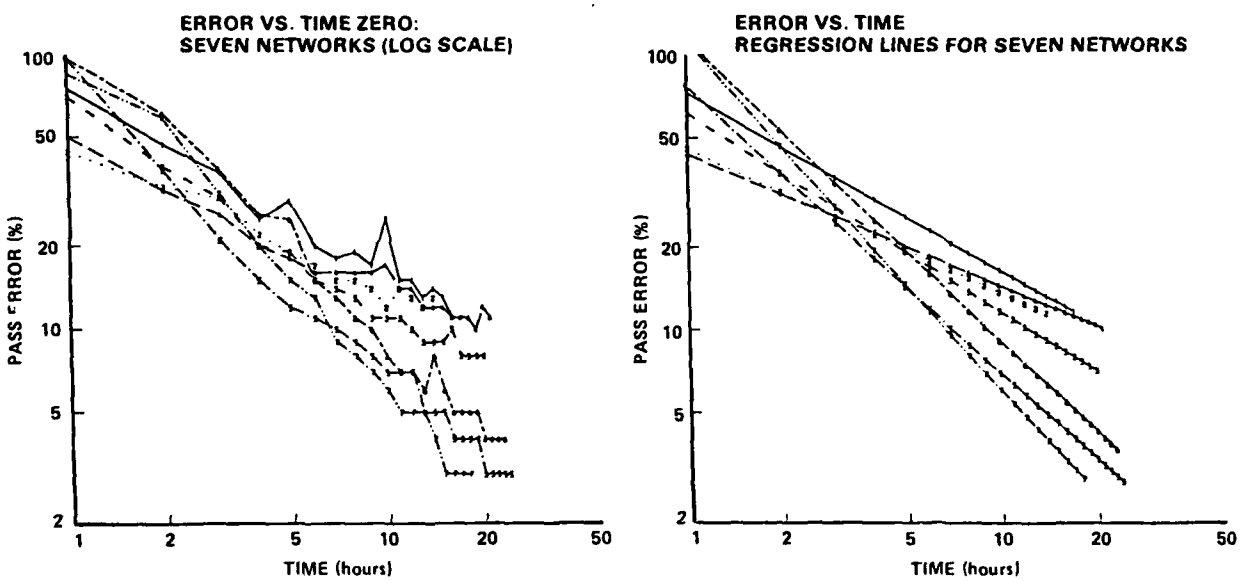

Fig. 2. Training Error vs. Time (Seven Networks).

As another initial experiment, we trained a network on one person's voice and then input a different person's voice through the network. What would the output sound like—the trainer's voice or the input voice? The output sounded like the input voice, not the trainer's voice. This was interesting, since it suggested the network was learning general features of speech, not unique features of a particular voice. Yet, our current results, described below, suggest the network may be trying to learn unique features as the network's hidden layer size decreases.

To determine if the network learned features relative to the training data, we listened to the reconstructed speech. Results are presented below.

OBSERVATIONS

Although it is too early to make firm conclusions, some specific observations concerning the initial and current results are provided below.

TRAINING RATE

The back-propagation network follows a power-law learning function (see Fig. 2).

NETWORK ERROR AND INVERSE PROCESSING

(Refer to Table III.) If the network could be trained to zero error, System Error would equal Processing Error for the Training Speaker. For Raw Speech, the network Training Error (Table II) is equal to the System Error, as expected. In the case of the FFT data, however, System Errors for the Training Speaker are much greater (2 to 5 times) than the Training Error. Therefore, the error introduced by the network is not additive and is *magnified* by the inverse speech processing.

PHASE INFORMATION AND THE NETWORK

Although the network can train in reasonable time using both FFT Power Spectra and Phase information (Table II), it has a difficult time generalizing phase information for speech reconstruction, resulting in larger System Errors (Table III).

NETWORK EFFICIENCY

Comparing the size **64-20-64** network with the size **64-10-64** network (Table III), note that the network with the smaller hidden layer causes less than a 3-percent increase in System Error. This is interesting since the smaller network has only one-half of the weights with which to store its mapping information.

COMPRESSION

As shown in Table III, both systems (raw and FFT) do an acceptable mapping of input speech to input speech, with a data compression ratio as high as 6.4 to 1.0.

MAPPING INPUT SPEECH TO INPUT SPEECH

(Refer to Table III.) For Raw Speech and FFT Power Spectra, a network trained to just under 10 percent error is able to generate speech features that result in an overall System Error of less than 50 percent. This shows that the network does in fact learn speech features. We would expect even lower System Errors if the network could be trained to a much lower Training Error (perhaps 1 percent).

MAPPING INPUT SPEECH TO TRAINING SPEECH

We wanted to see if the network maps a Test Speaker's voice to the Trained Speaker's voice. This is not clear from the data, so subjective listening tests were performed on the reconstructed speech.

In the case of Raw Speech, neither the identity of the Test Speaker nor the Training Speaker is discernible due to an increase in noise resulting from the System Error. However, the speech is intelligible, indicating that the network does learn speech features.

In the case of FFT Power Spectra, the identity of the Test Speaker is apparent in the reconstructed speech. This is not surprising, however, since the Test Speaker's original phase information was used to reconstruct the speech signal. Even though the System Errors are greater than in the Raw Speech case, the speech signal is highly intelligible.

In the case of FFT Power and Phase, noise in the reconstructed speech prevented any attempt to identify the speaker, and intelligibility was very low.

CURRENT CONCLUSIONS

From the above observations, we conclude the following: (1) Inverse speech processing magnifies errors introduced by the network. Therefore, the network must be trained to very low errors in a practical system. (2) The back-propagation network can compress speech effectively. In our case, 6.4-to-1. Further compression may be possible with larger networks. (3) Phase information contained in speech signals is difficult for a network to learn. It is desirable to parameterize phase information into a form more manageable by the network. (4) The network has difficulty learning significant representations of speech *which are unique to a training speaker's voice.* This is not totally surprising, since the back-propagation network is an *interpolative-associative* memory. If the network did learn features of the training speaker, they would not necessarily be present in the output of the network since it interpolates an input with its stored knowledge.

Our results, although informative, would be more useful if we could devise a speech processing system that allows the network to learn features with a more acceptable training error (1 or 2 percent). Our experience shows this could be possible if (1) more processing power was available to decrease training time, and (2) larger networks could be used. Both of these issues can be addressed by modifying the gradient descent algorithms used in training the networks, as well as running our experiments on faster machines. Additionally, other network paradigms, such as *acretive-associative* memories may be useful in attempting to map one person's voice to another.

FUTURE STUDY

The remainder of our research for this year will be to apply LPC speech data to the back-propagation network. The results of this work will be available by late summer of this year.

Future work will include study of acretive-associative networks paradigms for possible use in speech processing.

REFERENCES

1. Elman, J.L., & Zipser, D. *Learning the Hidden Structure of Speech.* University of California at San Diego, Institute for Cognitive Science. 1987.
2. Luse, Stephen A. "Neural Networks for Speech Applications," *Speech Technology Magazine.* Oct., Nov. 1987.
3. Sejnowski, T.J., & Rosenberg, C.R. *NET-talk: A parallel network that learns to read aloud.* Johns Hopkins University Department of Electrical Engineering and Computer Science Technical Report 86/01. 1986.
4. Klatt, Dennis H. "The Problem of Variability in Speech Recognition and in Models of Speech Perception," in *Invariance and Variability in Speech Processes,* ed. Joseph S. Perkell & Dennis H. Klatt. 1986.
5. Lea, Wayne A. *Trends in Speech Recognition,* Prentice-Hall. 1980.
6. Flanagan, J.L. *Speech Analysis, Synthesis, and Perception,* 2nd ed. 1983.