Neural network adaptive wavelets for signal representation and classification

Harold H. Szu
Brian Telfer
Department of the Navy
Naval Surface Warfare Center
Code R44
10901 New Hampshire Avenue
Silver Spring, Maryland 20903-5000

Shubha Kadambe
A. I. duPont Institute
Applied Science and Engineering Laboratories
Wilmington, Delaware 19899

Abstract. Methods are presented for adaptively generating wavelet templates for signal representation and classification using neural networks. Different network structures and energy functions are necessary and are given for representation and classification. The idea is introduced of a "super-wavelet," a linear combination of wavelets that itself is treated as a wavelet. The super-wavelet allows the shape of the wavelet to adapt to a particular problem, which goes beyond adapting parameters of a fixed-shape wavelet. Simulations are given for 1-D signals, with the concepts extendable to imagery. Ideas are discussed for applying the concepts in the paper to phoneme and speaker recognition.

Subject terms: wavelet transforms; classification; feature selection; neural networks; phoneme recognition; signal approximation; signal representation; speaker recognition.

Optical Engineering 31(9), 1907-1916 (September 1992).

1 Introduction

Wavelets show promise for both signal (or image) representation and classification. Signal representation using wavelets has received by far the most attention.[1-6] Representation and classification both can be viewed as feature extraction problems in which the goal is to find a set of daughter wavelets (dilations and shifts of a mother wavelet) that either best represent the signal or best separate various signal classes in the resulting feature space. However, the best set of wavelets for representation will not necessarily be the same as the best set for classification, or vice versa. This is because representation emphasizes the humps of a distribution, while classification emphasizes the overlapping tails, which tend to be close to the decision boundaries.

We present examples of how wavelets can be adaptively computed for representation and classification. By "adaptive," we mean that either the wavelet parameters or the wavelet shape are iteratively computed to minimize an energy function for a specific application. This differs from most previous applications that test a large fixed set of wavelets and then discard the ones that contribute least. In addition, a new concept of a "superposition-wavelet," or in short, "super-wavelet," is introduced. The super-wavelet is a linear combination of adaptive wavelets that is itself treated as a wavelet, in that dilations of a super-wavelet handle scale changes in a signal. The introduction of a super-wavelet means that the fundamental shape of the wavelet can be adapted to particular applications, rather than just the parameters of a fixed-shape wavelet.

As noted above, wavelets have been rarely applied to classification. One paper has considered this,[7] but has used human-selected rather than adaptive wavelets. We now review major approaches that have been used for representation. Most approaches, e.g., Ref. 1 and 2, use a fixed mother wavelet with varying shift and dilation parameters. Clearly, higher compression ratios can be obtained by choosing the mother wavelet that best fits the data. Coifman and Wickerhauser[3] find the best (in terms of minimum entropy) mother wavelets from a library of possible mother wavelets for compressing a signal. They have produced a fast algorithm ($N \log N$, where N is the signal length) for doing so. One would expect that higher compression ratios can be obtained by adaptively computing the wavelet, rather than selecting one from a fixed library, although this is likely to be at the cost of greater computation time. Tewfik, Singha, and Jorgensen[4] propose a complex but efficient approach for computing the best orthogonal mother wavelet from a scaling function. Several methods that combine neural networks and wavelets have been considered. Daugman[1] uses a neural network to learn the best set of coefficients for approximating an image with a set of Gabor wavelets. Pati and Krishnaprasad[5] also find the best coefficients for a wavelet expansion using a neural network. However, a more adaptive approach is to learn the wavelet parameters using neural networks, as Zhang and Benveniste[6] have done for approximating a function. Our approach is similar to Ref. 6 but differs in important aspects, namely, the wavelet function, the approximation function, the learning algorithm, and the super-wavelet concept. We also address classification as well as representation.

Throughout this paper, we use examples based on speech signals. That is because we feel this adaptive wavelet approach has great potential for both speech and machine-made sounds (as well as images). However, the principal purpose of the paper is to demonstrate the concept of adaptive wavelets, and not to actually solve any aspect of the speech problem, which remains for future work.

Paper WT-011 received April 20, 1992; revised manuscript received June 4, 1992; accepted for publication June 8, 1992.

Section 2 formulates the problem and the different neural network structures used for representation and classification. Simulation examples are given in Sec. 3, which also discusses the super-wavelet concept. Section 4 discusses how this approach might apply to phoneme and speaker recognition. Section 5 offers conclusions. An appendix provides physiological insights into speech characteristics.

2 Formulation

A network and formulation for signal representation is offered first and is followed by signal classification.

2.1 *Representation*

A signal $s(t)$ can be approximated by daughters of a mother wavelet $h(t)$ according to

$$\hat{s}(t)=\sum_{k=1}^{K} w_k h\left(\frac{t-b_k}{a_k}\right) , \tag{1}$$

where the w_k, b_k, and a_k are the weight coefficients, shifts, and dilations for each daughter wavelet. This approximation can be expressed as the neural network of Fig. 1, which contains wavelet nonlinearities in the artificial neurons rather than the standard sigmoidal nonlinearities. This architecture is similar to a radial basis function (RBF) neural network[31,32] because symmetric wavelets form a family of RBFs specified by the dilation parameter. The network parameters w_k, b_k, and a_k can be optimized by minimizing an energy function. We employ the least-mean-squares (LMS) energy for signal representation.

$$E=\frac{1}{2}\sum_{t=1}^{T}[s(t)-\hat{s}(t)]^2 . \tag{2}$$

A simple extension of Eq. (2) would be to produce an approximation over multiple realizations of a particular waveform to reduce noise and extract commonality. Adopting the mother wavelet

$$h(t)=\cos(1.75t)\exp(-t^2/2) \tag{3}$$

and letting $t'=(t-b_k)/a_k$, the gradients of E [Eq. (2)] are

$$g(w)_k=\frac{\partial E}{\partial w_k}=-\sum_{t=1}^{T}[s(t)-\hat{s}(t)]\cos(1.75t')\exp(-t'^2/2) , \tag{4}$$

$$g(b)_k=\frac{\partial E}{\partial b_k}=-\sum_{t=1}^{T}[s(t)-\hat{s}(t)]w_k[1.75\sin(1.75t') \times\exp(-t'^2/2)/a_k+\cos(1.75t')\exp(-t'^2/2)t'/a_k] , \tag{5}$$

$$g(a)_k=\frac{\partial E}{\partial a_k}=-\sum_{t=1}^{T}[s(t)-\hat{s}(t)]w_k[1.75\sin(1.75t') \times\exp(-t'^2/2)t'/a_k+\cos(1.75t') \times\exp(-t'^2/2)t'^2/a_k]=t'g(b)_k . \tag{6}$$

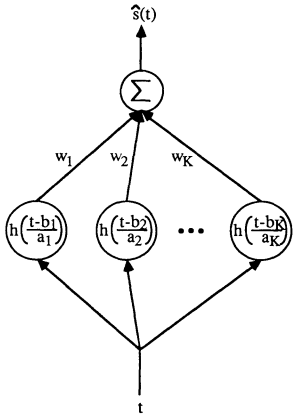


Fig. 1 Example neural network architecture for wavelet signal approximation, where the time value t feeds into the K nodes with wavelet nonlinearities.

We use a conjugate gradient method[8] to minimize E. Forming the column vectors $\mathbf{g}(\mathbf{w})$ and $\mathbf{w}$ from the elements $g(w)_k$ and w_k, the i'th iteration for minimizing E with respect to $\mathbf{w}$ proceeds according to the following two steps [$\mathbf{s}(\mathbf{w})$ is the search direction as a function of $\mathbf{w}$]:

1. if k is multiple of n then $\mathbf{s}(\mathbf{w})^i=-\mathbf{g}(\mathbf{w})^i$ (7)

$$\text{else } \mathbf{s}(\mathbf{w})^i=-\mathbf{g}(\mathbf{w})^i+\frac{\mathbf{g}(\mathbf{w})^{iT}\mathbf{g}(\mathbf{w})^i}{\mathbf{g}(\mathbf{w})^{(i-1)T}\mathbf{g}(\mathbf{w})^{(i-1)}}\mathbf{s}(\mathbf{w})^{i-1} , \tag{8}$$

2. $\mathbf{w}^{i+1}=\mathbf{w}^i+\alpha_w^i\mathbf{s}(\mathbf{w})^i$. (9)

Step 1 computes a search direction $\mathbf{s}$ at iteration i. Step 2 computes the new weight vector using a variable step-size α_w. At each iteration, steps 1 and 2 are computed for the representation parameter vectors $\mathbf{w}$, $\mathbf{a}$, and $\mathbf{b}$. It is preferable to perform a line search to find the best step-size (this can greatly reduce the number of iterations needed for convergence), but to demonstrate the concept of adaptive wavelets we use fixed step-sizes for simplicity.

2.2 *Classification*

Wavelets appear promising as a feature space for classification. The extraction of features in this case are the vector inner products of a set of wavelets with the input signal. These features can then be input to a classifier. A major issue is which wavelets to select. As an example of an adaptive solution to this problem, we consider the combined classifier and wavelet feature detector given by

$$v_n=\sigma(u_n)=\sigma\left[\sum_{k=1}^{K}w_k\sum_{t=1}^{T}s_n(t)h\left(\frac{t-b_k}{a_k}\right)\right] , \tag{10}$$

where v_n is the output for the n'th training vector $s_n(t)$ and $\sigma(z)=1/[1+\exp(-z)]$, a sigmoidal function. This classifier can be depicted as the neural network of Fig. 2, which uses wavelet weights rather than the wavelet nonlinearities of the representation network in Fig. 1. The lower part of Fig. 2 produces inner products of the signal and wavelets, with the first wavelet on the left and the K'th wavelet on the

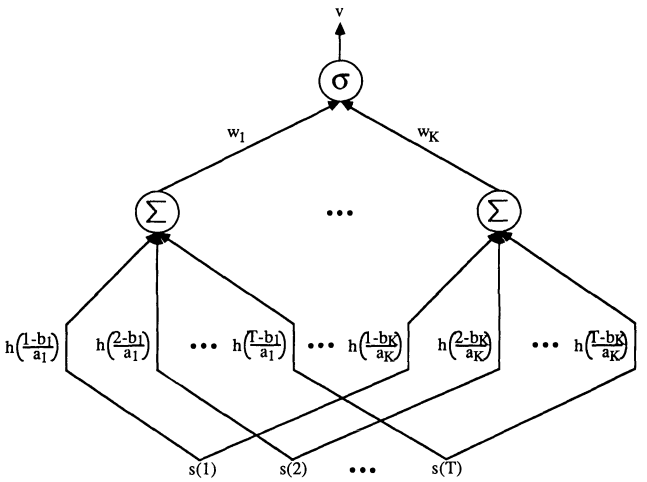


Fig. 2 Example neural network architecture for classifier with wavelet features (after synthesis all weights compress to single layer because of linearities).

right. Figure 2 shows two layers of weights, but once the network is synthesized, the two layers collapse into one because a nonlinearity does not exist between layers. The classification parameters w_k, b_k, and a_k can be optimized by minimizing, e.g., for two well-separated classes,

$$E=\frac{1}{2}\sum_{n=1}^{N}(d_n-v_n)^2 \ , \tag{11}$$

where d_n is the desired classifier output for $s_n(t)$. We set $d_n=1$ for one class and $d_n=0$ for the other. Extensions or other approaches are certainly possible. For example, a more complex multilayer network or a network with multiple output elements to handle more than two classes could be adopted instead of the classifier in Eq. (10). More than one mother wavelet could be included in Eq. (10) [e.g., in addition to $h(t)$, one could include another mother $g(t)$]. Also, a different measure could be used to determine the optimal features, such as the Fisher ratio[9] or minimax criterion, which minimizes the difference between the intraclass variation of the training vectors and maximizes the interclass variation,[10] or a minimum-misclassification-error criterion for overlapping classes.[11] However, the approach of Eqs. (10) and (11) suffices to demonstrate the concept of adaptive wavelet feature generation. Employing the wavelet of Eq. (3) and letting $t'=(t-b_k)/a_k$, $\sigma'(u)=\partial\sigma(u)/\partial u=\sigma(u)[1-\sigma(u)]$, the gradients of E [Eq. (11)] are

$$g(w)_k=-\sum_{n=1}^{N}\sum_{t=1}^{T}(d_n-v_n)\sigma'(u_n)s_n(t)\cos(1.75t')\times\exp(-t'^2/2) \ , \tag{12}$$

$$g(b)_k=-\sum_{n=1}^{N}\sum_{t=1}^{T}(d_n-v_n)\sigma'(u_n)s_n(t)w_k[1.75\sin(1.75t')\times\exp(-t'^2/2)/a_k+\cos(1.75t')\exp(-t'^2/2)t'/a_k] \ , \tag{13}$$

$$g(a)_k=-\sum_{n=1}^{N}\sum_{t=1}^{T}(d_n-v_n)\sigma'(u_n)s_n(t)w_k\times[1.75\sin(1.75t')\exp(-t'^2/2)t'/a_k+\cos(1.75t')\exp(-t'^2/2)t'^2/a_k]=t'g(b)_k \ , \tag{14}$$

Conjugate gradient descent is used to minimize Eq. (11), as described in Sec. 2.1. We have proposed different network structures and energy functions for representation and classification. We next present results run on sample data for the formulations in this section.

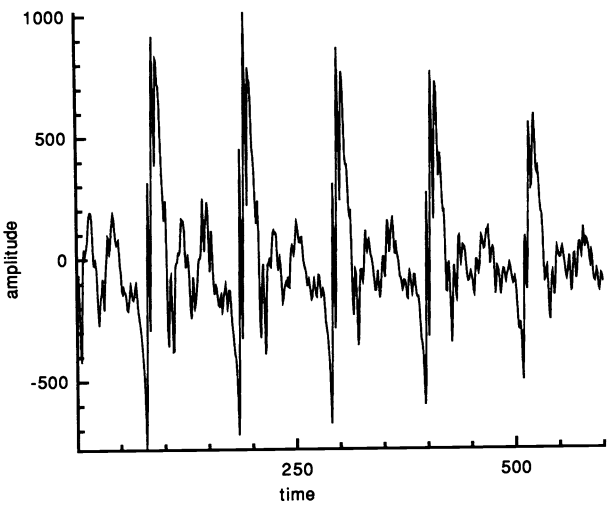
3 Simulations

Simulation results for signal representation are offered first, followed by signal classification.

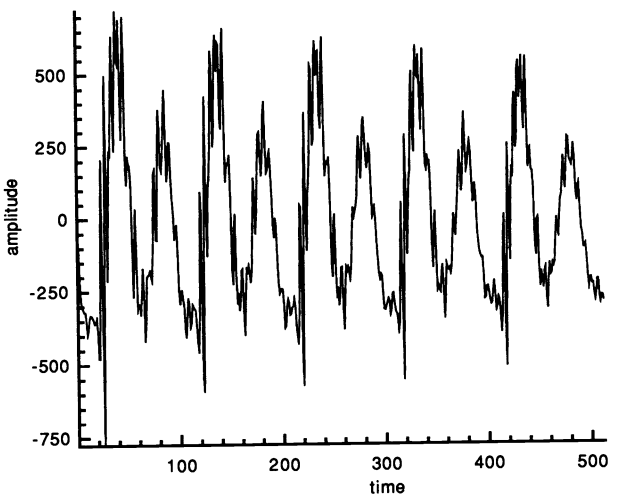
3.1 *Representation*

To demonstrate how adaptive wavelets can approximate functions, we consider three phonemes, "a," "e," and "i," that were extracted from speech signals and which are shown in Fig. 3. (These are long vowels spoken in isolation.) In this section, we consider the phonemes as generic signals to demonstrate the neural network's operation. In Sec. 4, we consider how this method could be applied to phoneme and speaker recognition in future work. Note that each phoneme in Fig. 3 is periodic. We approximate only a single period of each. The solid lines in Fig. 4 show the extracted periods, where the signal on either side of the period has been windowed with a Gaussian falloff with a standard deviation of two pixels. The dashed lines in Fig. 4 show the wavelet approximations, which we now describe. The wavelet in Eq. (3) with $a=8$ is shown in Fig. 5. To determine the number and initial placement for each wavelet, we convolve this wavelet with the signal and place a mother wavelet at each location where a peak occurs. The number of wavelets selected for "a," "e," and "i" was 11, 6, and 14. All weights and dilations were initialized to 0 and 8, respectively. The gradient descent algorithm was run for 500 iterations (batch mode) with $\alpha_w=10^{-2}$, $\alpha_a=\alpha_b=10^{-7}$, and a restart cycle of $n=10$. Letting $\mathbf{e}=\mathbf{s}-\hat{\mathbf{s}}$ (where the boldface denotes column vectors containing the time samples of the signal, etc.), we measure a normalized approximation error $\mathbf{e}^T\mathbf{e}/\mathbf{s}^T\mathbf{s}$, so that the error between $\mathbf{s}$ and an all-zero vector $\mathbf{0}$ equals 1. The approximation errors for "a," "e," and "i" are 0.210, 0.231, and 0.083. Tables 1, 2, and 3 show the final parameters for each approximation. Note that the dilations range over an order of magnitude, with the smallest being 0.207 times the initial value of 8, and the largest being twice the initial value. The maximum change between an initial and final shift is 9.8, or roughly half of the initial spacing between shifts. Thus, the neural network has adaptively created a wide range of daughter wavelets and has produced good approximations.

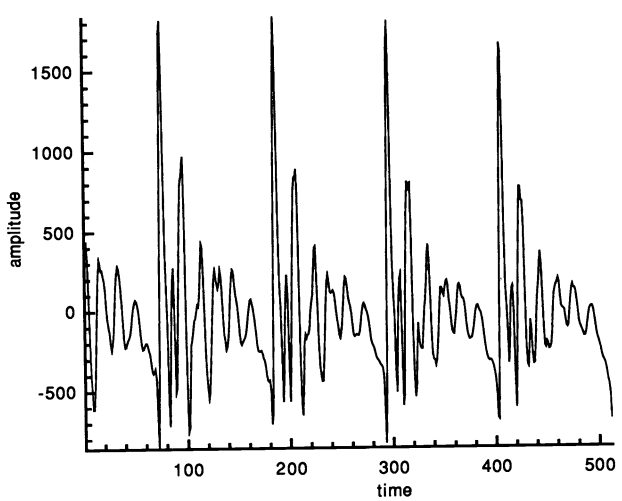
The purpose of these simulations is to show the potential of neural networks for adaptively creating a wavelet approximation, not to produce an efficient production code for doing so. The program required three CPU minutes for approximating the single period of "i," which is represented by 14 wavelets and required proportionately less for the other two phonemes. The speed of neural network synthesis could be dramatically improved by (1) incorporating



(a)

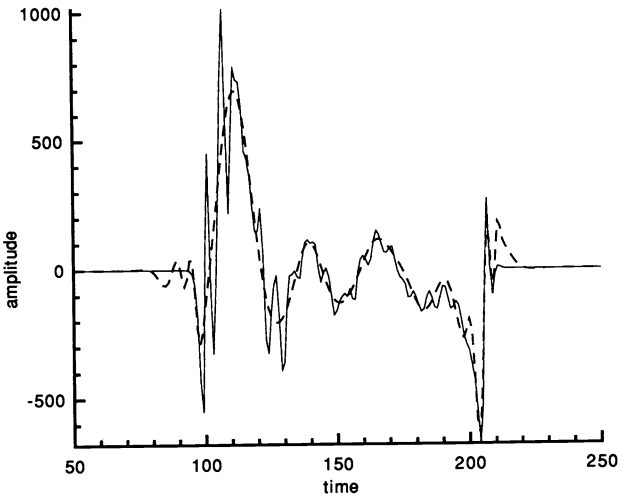


(b)

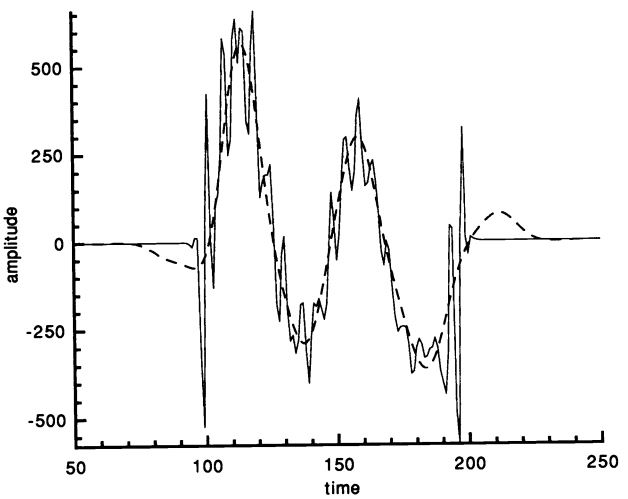


(c)

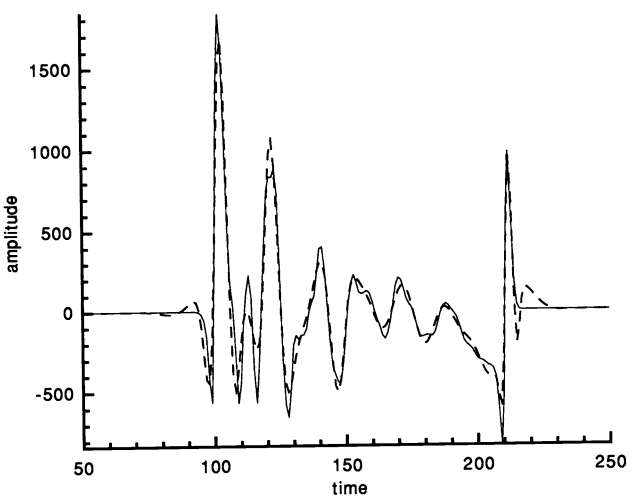
Fig. 3 Phonemes: (a) "a," (b) "e," and (c) "i."



(a)



(b)



(c)

Fig. 4 Single periods of phonemes extracted from Fig. 3 signals (solid lines) and adaptive wavelet approximations (dashed lines): (a) "a," (b) "e," and (c) "i."

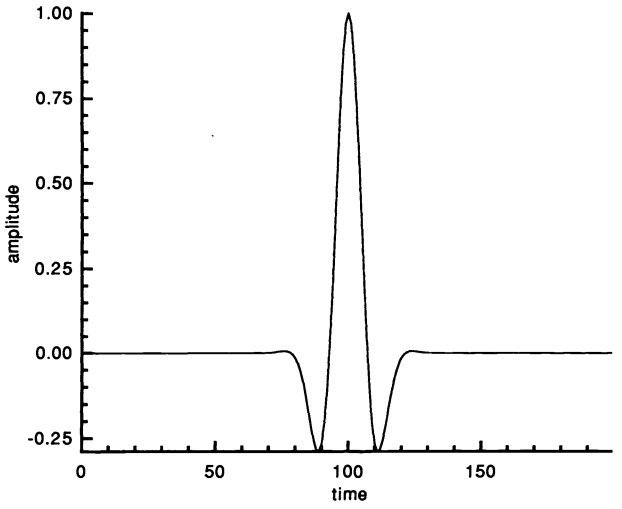


Fig. 5 Wavelet given by Eq. (3) with $a=8$.

Table 1 Weights, dilations, and shifts for adaptive wavelet approximation of "a."

Wavelet Number	Weights w	Dilations a	Shifts b	
			Initial	Final
1	165	4.72	86	90.7
2	238	2.42	98	94.9
3	700	11.0	111	111
4	15.0	8.89	127	128
5	85.9	6.73	140	139
6	-110	8.47	152	151
7	49.8	6.39	166	165
8	-220	10.7	181	185
9	-394	8.22	190	200
10	-468	2.66	201	204
11	-259	1.66	214	209

a line search that computes the best step-sizes α at each iteration, (2) including a stopping criterion, (3) creating better initial values for the weights and dilations, and (4) running the algorithm on specialized neural network hardware.

Note that the wavelet approximation of each signal describes a super-wavelet, which is a linear combination of wavelets that itself can be treated as a wavelet. Dilations of the super-wavelet could represent the same period of speech spoken at different speeds. Section 4 provides a more detailed discussion.

3.2 *Classification*

To generate a simple training set for demonstration purposes, we segmented three single-period training vectors from each of the "a," "e," and "i" signals. The length of each period was adjusted to be identical, and all nine signals were normalized to unit norm. These training vectors are plotted in Figs. 6(a) through 6(c). Because the three classes are quite different, they pose a simple recognition problem. To make the problem more challenging, 10 additional vectors for each class were synthesized by adding Gaussian noise with $\sigma=0.2$ (SNR $=-6$ dB) to the first period extracted from each class. A representative noisy training vector "a" is plotted in Fig. 6(d). Thus the training set contained 39 vectors. A test set was not used because we are simply demonstrating the concept of adaptive wavelet feature selection and not testing a real application. As in Sec. 3.1, these signals are treated as generic and static (we compute

Table 2 Weights, dilations, and shifts for adaptive wavelet approximation of "e."

Wavelet Number	Weights w	Dilations a	Shifts b	
			Initial	Final
1	96.9	10.1	98	96.4
2	588	12.4	114	114
3	-193	9.32	136	139
4	182	9.24	158	158
5	-372	16.0	186	184
6	-65.3	9.81	203	203

Table 3 Weights, dilations, and shifts for adaptive wavelet approximation of "i."

Wavelet Number	Weights w	Dilations a	Shifts b	
			Initial	Final
1	62.6	7.86	92	93.9
2	1720	3.68	103	103
3	35.3	7.05	112	113
4	1070	4.13	121	122
5	-173	9.09	131	130
6	138	7.89	139	140
7	-511	4.34	148	147
8	136	7.28	157	157
9	-104	7.48	164	164
10	104	7.39	172	172
11	-271	7.29	181	180
12	-295	7.84	190	193
13	-492	9.29	203	205
14	1050	2.20	215	212

only vector inner products, not correlations), and this is not intended to be a test of phoneme recognition.

A two-class case was tested, with all "a" vectors forming one class (desired output of 1) and all "e" and "i" vectors forming the other (desired output of 0). We chose to use four wavelet features after empirically determining that number was sufficient to classify the data. The Eq. (3) wavelets were initialized to equal dilations $a_k=16$ ($k=1, \ldots, 4$), shifts evenly spaced across the signals, $b_1=62$, $b_2=87$, $b_3=112$, $b_4=137$, and zero-valued weights. Leaving the wavelets fixed at the initial values, we first minimized E [Eq. (11)] by adapting only the weights (150 iterations, $\alpha_w=0.1$, $\alpha_a=\alpha_b=0.0$, restart cycle of 10). This resulted in five classification errors, or a 13% misclassification rate. Minimizing with adaptive wavelet features (150 iterations, $\alpha_w=\alpha_a=\alpha_b=0.1$) reduced the classification errors to 1, or 2.5%. This demonstrates the point that adaptive wavelet features can produce much better classification rates than ad hoc fixed wavelet features. Table 4 gives the resulting parameters for each case and shows that for the adaptive case, the dilations and shifts changed from their initial values. Figure 7 shows the resulting wavelet features for each case. Because the classifier linearly combines the wavelet features, Fig. 7 plots the linear combination of the weighted wavelets. (Linear combination of wavelets has previously been used to form a detection filter,[7] but the wavelets were fixed and not adaptive.) Figure 7 clearly shows the differences between the fixed and adaptive wavelet features. In a real application, one would want to pick good initial values for the wavelet parameters and then optimize them as demonstrated in this paper. Good initial values are important to avoid local minima of energy functions such as Eq. (11). The combination of wavelets in Fig. 7(b) forms a super-wavelet meant for classification rather than representation.

Features rather than raw data are used for classification for several reasons: (1) reducing the dimension of data makes

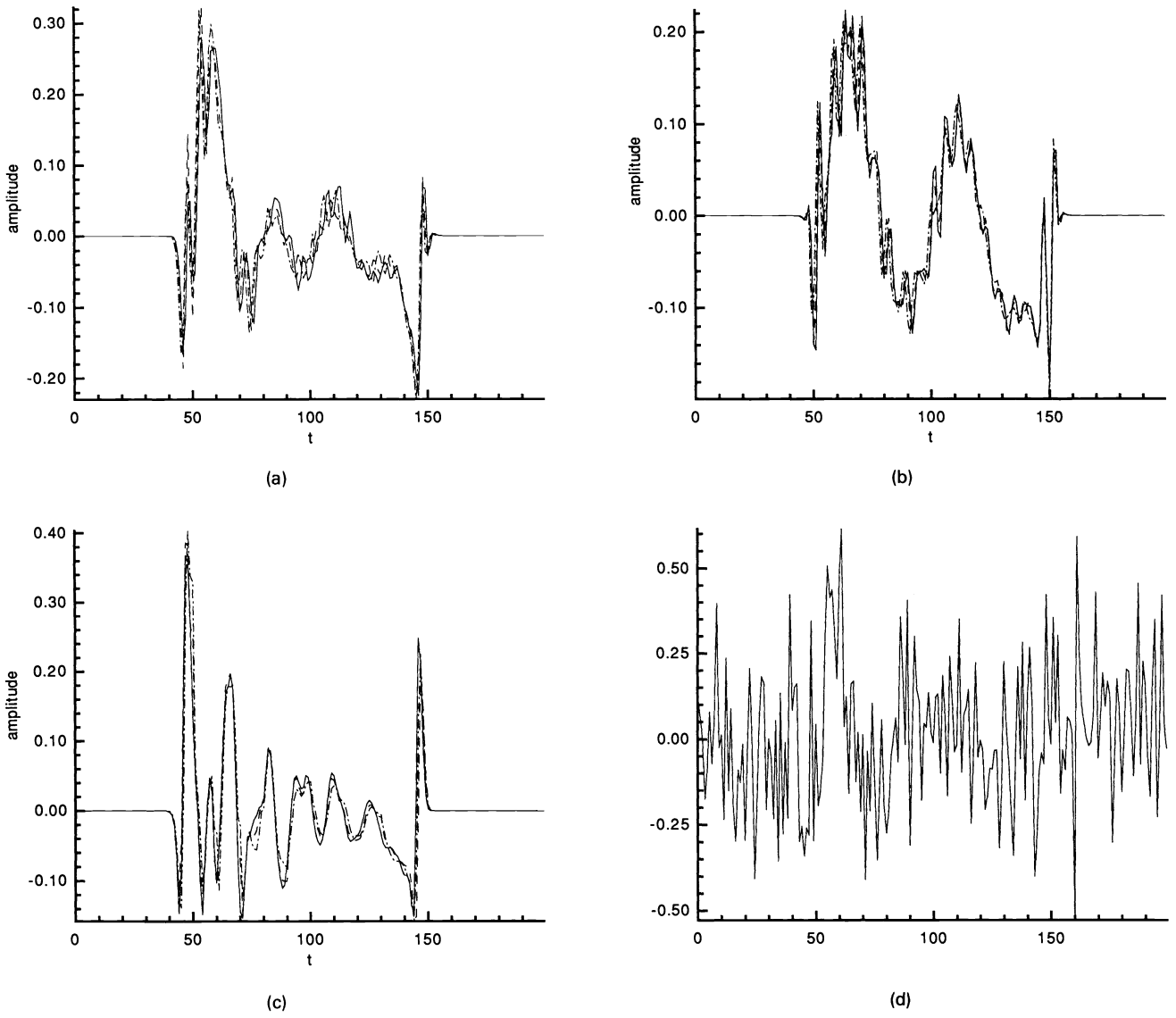


Fig. 6 Training vectors: three single periods (solid, dashed, and dashed-dot lines) from (a) "a," (b) "e," (c) "i," and (d) noisy "a" training vector (ten noisy training vectors used for each class).

Table 4 Weights, dilations, and shifts for fixed and adaptive wavelet features for classification.

Wavelet Number	Fixed Features			Adaptive Features		
	w	a	b	w	a	b
1	2.59	16.0	62.0	3.89	9.8	59.7
2	2.73	16.0	87.0	3.59	13.7	84.7
3	0.24	16.0	112.0	0.76	15.7	111.9
4	2.40	16.0	137.0	4.30	13.5	133.4

the problem more overdetermined by the training set and therefore can increase the classification rate, (2) reducing the dimension of the data speeds up training, and (3) features can incorporate invariances such as scale, translation, etc., to avoid impractically large training sets. The adaptive wavelet features we have demonstrated primarily address the first reason. The adaptive nature requires more computation time than fixed features, but this second reason is a relatively minor issue, since training is normally performed off-line. This paper has not addressed the third reason, which represents an important issue for future work. Invariances seem less important for speech signals than for images, since scale changes in speech can conceivably be handled by wavelet dilations, while classifying objects in images often requires rotation invariance.

Sections 2 and 3 clearly show that representation and classification significantly differ in terms of the network structure and the type of criterion that is optimized and in terms of the resulting wavelets. However, both approaches can be used for recognition, as described in Sec. 4.

4 Speech Case Study

To make the concepts presented in the previous sections more concrete, we consider how these ideas might apply to phoneme and speaker recognition. Implementation of these ideas to a particular application remains for future work.

American English speech is composed of 42 basic sounds, or phonemes.[12] The phonemes are broadly classified as being voiced, unvoiced, and mixed.[13] Voiced sounds are periodic

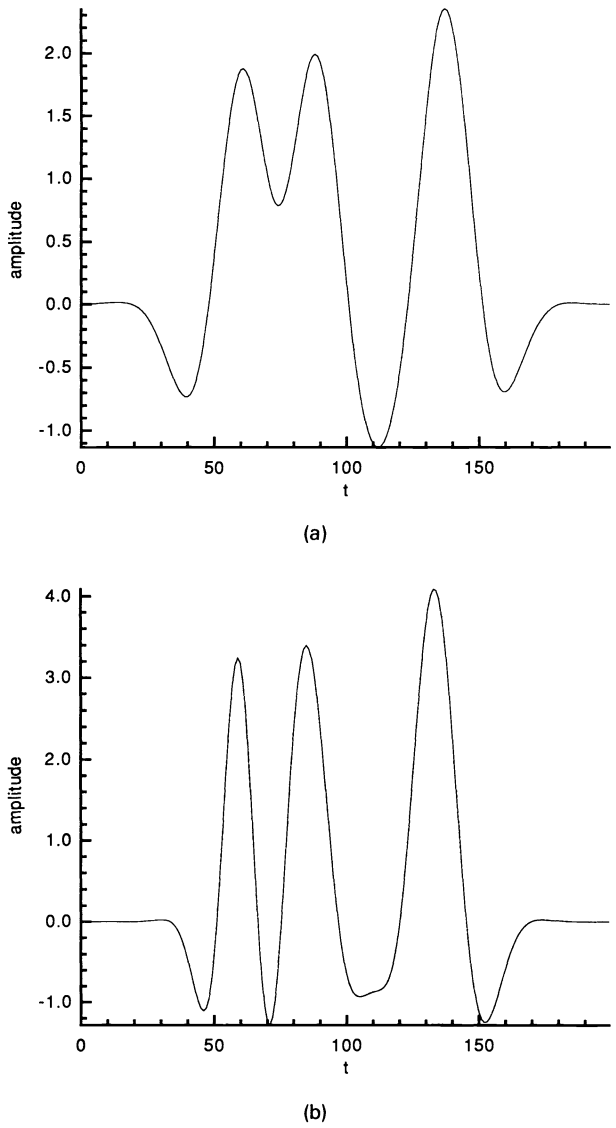


Fig. 7 Wavelet features: (a) fixed and (b) adaptive.

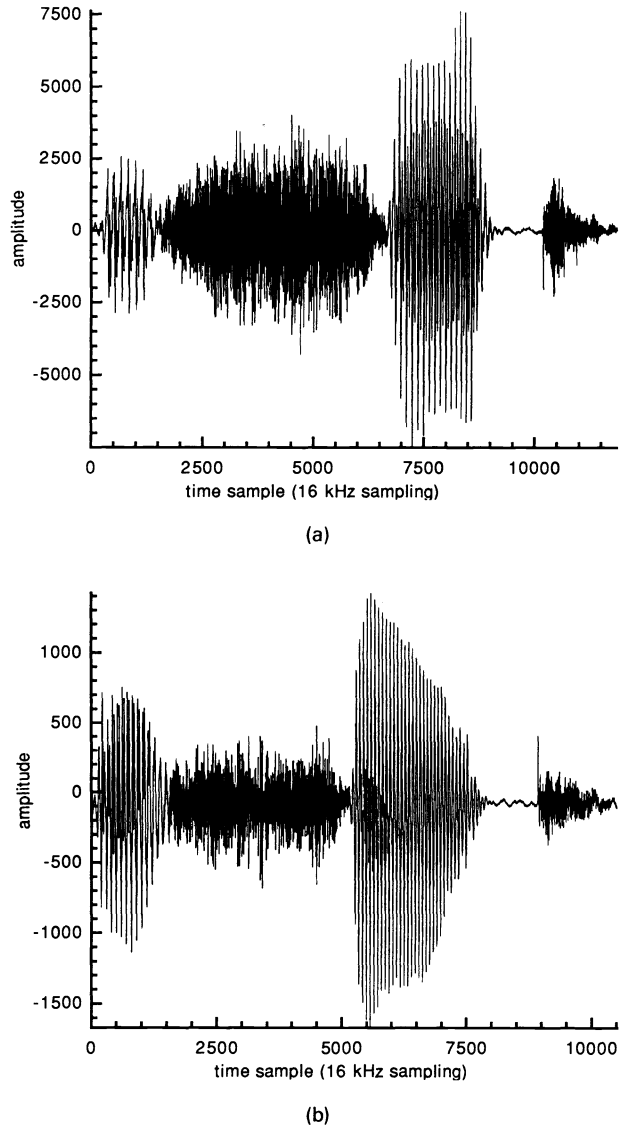


Fig. 8 American (a) male and (b) female speakers saying "the seat," extracted from conversational speech with a 16-kHz sampling rate.

or semiperiodic (e.g., the "a," "e," and "i" phonemes in Fig. 3). Unvoiced sounds are higher frequency and more noiselike. The waveform of a phoneme varies from phoneme to phoneme, from speaker to speaker, and from pronunciation to pronunciation for the same speaker. Figure 8 shows examples of speech using the words "the seat," in which the periodic nature of the voiced sounds are visible and the high-frequency noiselike quality of unvoiced sounds can be seen in the "s" and "t." Also, the two signals give an idea of how speech differs in frequency and envelopes between two speakers, particularly between male and female speakers. More details on the physiology of speech characteristics and speaker differences are given in Sec. 6. We make use of the waveform variability of phonemes to suggest a phoneme recognition system and the variability of the same phoneme from speaker to speaker to suggest a speaker recognition system using adaptive super-wavelets. Sections 4.1 and 4.2 discuss a phoneme recognition and a speaker recognition system.

4.1 *Phoneme Recognition*

Phoneme recognition systems are used in automatic speech recognition systems and related phoneme generators are used in speech synthesis systems. Several phoneme recognizers have been developed.[14–18] These systems exploit the variations in the phonemes' spectra by computing the spectrum of a small segment of a speech signal and then computing the mel or bark scale coefficients from the power spectrum. These features are classified by classical methods[14] and neural network approaches.[15–18]

Adaptive wavelets offer two potential approaches. First, a super-wavelet could be generated for each phoneme using the function representation method of Sec. 2.1, optimized over multiple speakers. The super-wavelet might be fashioned to represent several periods of a phoneme to improve the SNR. The set of super-wavelets then forms a bank of filters that can be correlated with a speech signal. The correlation peaks would identify the phonemes. Dilated ver-

sions of the super-wavelets could be used to identify speech at different speeds. This idea is shown conceptually in Fig. 9, which plots correlations of the super-wavelets in Fig. 4 (normalized to unit norm) with the full phoneme signals in Fig. 3. The correlation peaks clearly indicate the occurence of each period and the type of phoneme. For example, Fig. 9(a) plots the correlations of the three super-wavelets with the "a" phoneme and the highest correlation peaks are in the solid-line plot produced by the "a" super-wavelet. The correlation peaks decrease over time because the signal strength is decreasing. Clearly, the local signal strength must be taken into account in such an approach. This is a simplistic example, in that Fig. 9 is testing on the training data (at least for the first period of the signal) and the wavelet approximations have not been produced over multiple phoneme realizations, but Fig. 9 is only meant to show the concept. This approach resembles template matching, except that the super-wavelet is produced from multiple phoneme realizations and wavelet dilations handle speed changes.

In the second approach, a classifier with adaptive wavelet features as in Sec. 2.2 could be used to identify phonemes. This is similar to the classifiers described above, except that instead of features taken from a spectrogram, the adaptive wavelets generate wideband transient features that are tailored to the problem. By optimizing the features for the problem, fewer features should be needed and better classification rates could result. The adaptive wavelet approach seems best suited to the better characterized waveforms of voiced rather than unvoiced phonemes.

4.2 *Speaker Recognition*

Two main applications of speaker recognition are (1) verifying a person's identity prior to admitting him to a secured place or to a telephone transaction and (2) associating a person with a voice in police work.[19] Due to fewer applications of speaker recognition compared to speech recognition and lack of complete knowledge about which characteristics of a speech signal help in identifying a speaker, speaker recognition has received less emphasis.[20–24]

In general, automatic speaker recognizers exploit the variability in speech characteristics of different speakers caused by variations in the vocal cords and vocal tract. The differences in different speakers' vocal cords introduce variations in the pitch period (fundamental frequency of a speech signal), and differences in the vocal tract introduce variations in its resonant frequencies and, hence, variations in the waveform or spectrum of a phoneme.

The speaker recognizers developed so far can be broadly classified into text-dependent and text-independent systems. Text-dependent systems use a specially designed utterance, whereas text-independent systems operate on previously unknown speech utterances. The error rate is lower in the case of the text-dependent systems; however, text-independent systems are more flexible and foolproof. Hence, we consider a text-independent speaker recognition system.

Generally, text-independent speaker recognition systems use a feature set averaged over a long utterance for classification purposes.[25–27] The main disadvantage of long-term statistics is that they are often impractical for real-time text-independent applications. However, this problem can be overcome by using phonemes. Few phoneme-based speaker recognition systems have been developed.[24,28] One of these[24]

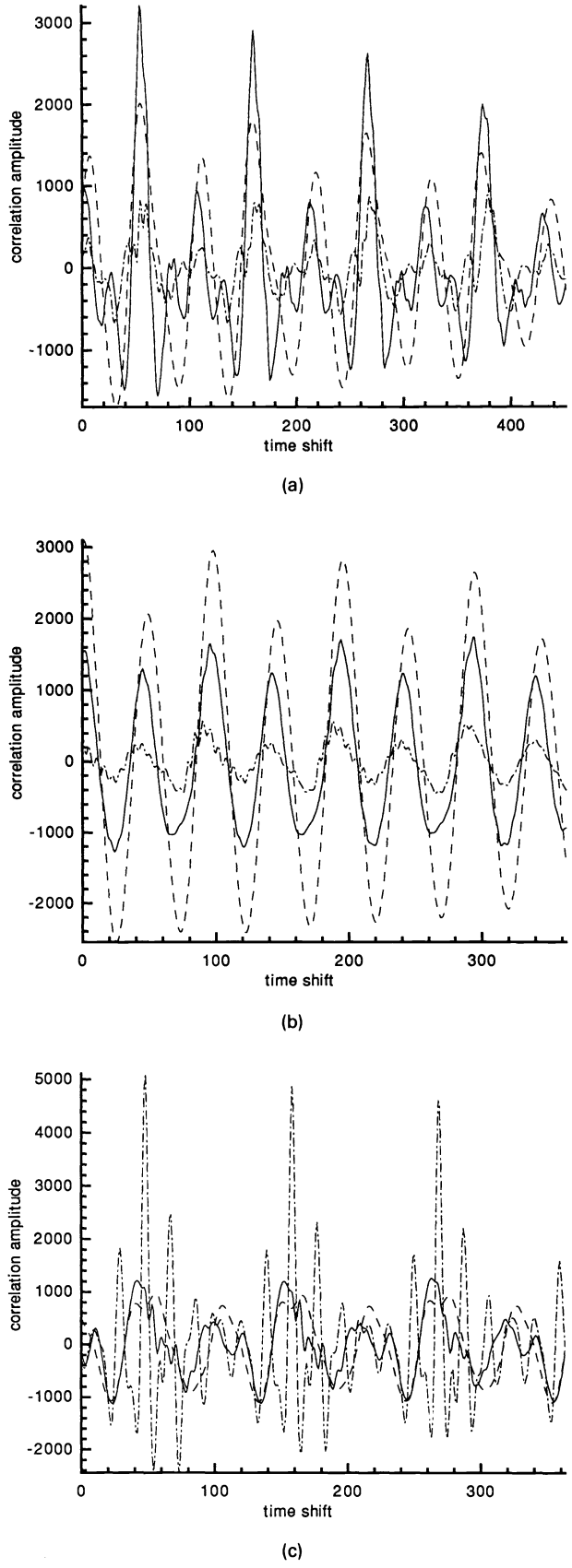


Fig. 9 Correlation of (a) "a," (b) "e," and (c) "i" phonemes in Fig. 3 with "a," "e," and "i" super-wavelets in Fig. 4, plotted with solid, dashed, and dashed-dot lines, respectively.

uses linear-predictive-coding (LPC) cepstral coefficients as features for a quadratic classifier.

The same adaptive wavelet approaches outlined in Sec. 4.1 can be applied to speaker recognition, except in this case the same phoneme from different speakers would form different classes. However, speaker recognition has the advantage that the spectral features of vowels (voiced phonemes) are most useful for speaker recognizers.[29] This is convenient for adaptive wavelets that can better capture the waveforms of voiced sounds than unvoiced.

5 Conclusion

Wavelets frequently have been applied to representation, but rarely to classification. We have shown how wavelets can be adaptively computed for either task, using different neural network structures and energy functions best suited for each. The new concept of a super-wavelet allows the wavelet shape to be adaptively computed for a particular problem, rather than only adaptively computing the parameters of a fixed-shape wavelet.

Our concern is primarily with classification rather than representation. For representation, orthogonal wavelets have proven very useful for efficient and fast data compression, e.g., Ref. 3. The adaptive wavelets we studied are not orthogonal, but we see this as less of an issue for classification where we are trying to find features that separate the classes rather than orthogonal features.

The concepts of adaptive wavelets have been demonstrated on 1-*D* signals, and a discussion was presented on how these concepts could apply to phoneme and speaker recognition. However, these concepts should also apply to images. In particular, the idea of using dilations of a super-wavelet to handle input scale changes applies to both 1-*D* signals and images. The idea of adaptively generating an optimal set of wavelet features seems like a powerful approach for both signals and images.

6 Appendix: Speech Characteristics

Speech characteristics arise from physiology. This appendix reviews physiological details to provide insight into interphoneme and interspeaker differences. The human speech production system consists of the lungs, trachea (windpipe), pharynx (throat cavity), and vocal tract (which includes the oral and nasal cavities).[30] Speech sounds are produced by the passage of forced air from the lungs through the trachea into the pharynx. The upper portion of the trachea contains a cartilaginous structure called the larynx. The larynx houses two liplike ligaments called the vocal cords. The slitlike opening between these two vocal cords is called the glottis. The vocal cords are held by arytenoid cartilage. This cartilage facilitates in adjusting the tension in the vocal cords. The air from the pharynx then passes through the oral or nasal cavity of the vocal tract depending on whether the velum (soft palate at the rear of the roof of the mouth) is closed or open.

A language can be described by a set of linguistic units called phonemes (distinct speech sounds).[19] For example, American English can be described by a set of 42 phonemes.[12] The nature of each phoneme varies based on the source of excitation (forced air), i.e., manner of articulation and the shape of the vocal tract, i.e., place of articulation. The shape of the vocal tract varies while producing various sounds based on the movements of articulatories such as the glottis, the pharynx, the velum, the jaw, the tongue, and the lips. The variations in the characteristics of a phoneme based on the shape of the vocal tract can be explained as follows: The vocal tract can be considered similar to an acoustic tube. The forced air from the lungs and the pharyngeal cavity causes the vocal tract to resonate, which modulates the sound waveform, and the resonant frequencies (formant frequencies) of the vocal tract vary depending on the length and shape of the vocal tract. The length of the vocal tract is fixed for a given speaker but varies from speaker to speaker.

Based on the source of excitation, phonemes can be broadly classified into voiced, unvoiced, and mixed voiced or mixed unvoiced sounds.[13] The voiced sounds are produced by the periodic or the semiperiodic vibrations of the vocal cords. The period of the vocal cords' vibrations depends on the mass and compliance of the vocal cords and the subglottal pressure (air pressure below the glottis). The unvoiced sounds are produced by a turbulent flow of air created by some constriction in the vocal tract. During the production of unvoiced sounds, the vocal cords are held apart and the glottis is fully open. The mixed sounds are produced by the abrupt release of air pressure built up due to closure at some point in the vocal tract. The abrupt release of air pressure provides transient excitation of the vocal tract. The transient excitation can be associated with or without the vocal cord vibrations producing mixed voiced or mixed unvoiced sounds.

Based on the place of articulation, the speech sounds can be classified into the following eight groups[19]:

1. *Labials:* If both lips are held together, the sound is called bilabia; if the lower lip is in contact with the upper teeth, the sound is called labio dental.
2. *Dental:* If the tongue tip or blade touches the edge or back of the upper incisor teeth, the sound is called dental.
3. *Alveolar:* If the tongue tip or blade approaches or touches the alveolar ridge (the ridge in the jaw where the teeth sockets are located) then the sound is called alveolar.
4. *Palatal:* If the tongue blade (dorsum) constricts with the hard palate or if the tongue tip curls, the sound is called palatal.
5. *Velar:* If the dorsum approaches the soft palate, the sound is called velar.
6. *Uvular:* If the tongue dorsum approaches the uvula, the sound is called uvular.
7. *Pharyngeal:* If the pharynx constricts, the sound is called pharyngeal.
8. *Glottal:* If the vocal cords are either close or constricted, the sound is called glottal.

From the above description, it is clear that the variety of body parts involved in producing speech create a rich variation in the spectral and the waveform nature of different phonemes.

Acknowledgments

The support of this research by the Naval Surface Warfare Center Dahlgren Division White Oak (NSWCDDWO) In-

dependent Research Program and an Office of Naval Research Young Navy Scientist Award is gratefully acknowledged.

References

1. J. Daugman, "Complete discrete 2-*D* Gabor transforms by neural networks for image analysis and compression," *IEEE Trans. Acoust., Speech, Signal Proc.* **36,** 1169–1179 (July 1988).
2. R. DeVore, B. Jawerth, and B.Lucier, "Image compression through wavelet transform coding," *IEEE Trans. Inf. Theory* **38,** 719–746 (March 1992).
3. R. Coifman and M. Wickerhauser, "Entropy based algorithms for best basis selection," *IEEE Trans. Inf. Theory* **38,** 713–718 (March 1992).
4. A Tewfik, D. Singha, and P. Jorgensen, "On the optimal choice of a wavelet for signal representation," *IEEE Trans. Inf. Theory* **38,** 747–765 (March 1992).
5. Y. Pati and P. Krishnaprasad, "Analysis and synthesis of feedforward neural networks using discrete affine wavelet transformations," Tech. Rep. SRC-TR-90-44, Univ. Maryland Systems Research Center (1991).
6. Q. Zhang and A. Benveniste, "Approximation by nonlinear wavelet networks," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing* **5,** 3417–3420 (May 1991).
7. D. Casasent, J.-S. Smokelin, and A. Ye, "Optical Gabor and wavelet transforms for scene analysis," *Proc SPIE* **1702** (April 1992).
8. R. Fletcher, *Practical Methods of Optimization,* John Wiley and Sons, New York (1987).
9. R. Duda and P. Hart, *Pattern Classification and Scene Analysis,* John Wiley and Sons, New York (1973).
10. H. Szu, "Neural networks based on Peano curves and hairy neurons," *Telematics Informatics* **7**(3/4), 403–430 (1990).
11. B. Telfer and H. Szu, "Implementing the minimum-misclassification-error energy function for target recognition," *Proc. IEEE Int. Joint Conf. Neural Networks—Baltimore* **4,** 214–219 (June 1992).
12. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals,* Prentice-Hall, New Jersey (1978).
13. J. W. Pickett, *The Sounds of Speech Communication: a Primer of Acoustic Phonetics and Speech Perception,* University Park Press, Baltimore (1980).
14. K. Tanaka, "A parametric representation and a clustering method for phoneme recognition—application to stops in a CV environment," *IEEE Trans. Acoust., Speech, Signal Proc.* **29,** 1117–1127 (December 1981).
15. A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. Acoustics, Speech, Signal Proc.* **37,** 328–339 (March 1989).
16. F. Greco, A. Paoloni, and G. Ravaioli, "A recurrent time-delay neural network for improved phoneme recognition," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* **1,** 81–84 (May 1991).
17. M. Nakamura, S. Tamura, and S. Sagayama, "Phoneme recognition by phoneme filter neural network," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing* **1,** 85–88 (May 1991).
18. J. Takami and S. Sagayama, "A pairwise discriminant approach to robust phoneme recognition by time delay neural networks," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* **1,** 89–92 (May 1991).
19. D. O'Shaughnessy, *Speech Communication Human and Machine,* Addison Wesley, New York (1990).
20. U. Goldstein, "Speaker-identifying features based on formant tracks," *J. Acoust. Soc. Am.* **59,** 176–182 (1976).
21. G. Doddington, "Speaker-recognition: identifying people from their voice," *Proc. IEEE* **73,** 1651–1664 (1985).
22. F. Soong, A. Rosenberg, L. Rabiner, and B. Juang, "A vector quantization approach to speaker recognition," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing,* pp. 387–390 (May 1985).
23. H. Hattori, "Text-independent speaker recognition using neural networks," *Proc. Int. Conf. Acoustics, Speech, and Signal Processing* **II,** 153–156 (March 1992).
24. M. Savic and J. Sorensen, "Phoneme based speaker verification," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing* **II,** 165–168 (March 1992).
25. H. Hollien and W. Majewski, "Speaker identification by long-term spectra under normal and distorted speech conditions," *J. Acoust. Soc. Am.* **62,** 975–980 (1977).
26. J. Markel and S. Davis, "Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base," *IEEE Trans. Acoustics, Speech, Signal Proc.* **27,** 74–82 (1979).
27. K. Li and G. Hughes, "Talker differences as they appear in correlation matrices of continuous speech spectra," *J. Acoust. Soc. Am.* **55,** 833–837 (1974).
28. A. Higgins and R. Wohlford, "A new method of text-independent speaker recognition," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing,* pp. 869–872 (1986).
29. F. Nolan, *The Phonetic Bases of Speaker Recognition,* Cambridge University Press, Cambridge (1983).
30. J. L. Flanagan, *Speech Analysis, Synthesis and Perception,* Springer-Verlag, 2nd expanded ed., New York (1972).
31. D. Broomhead and D. Lowe, "Multi-variable functional interpolation and adaptive networks," *Complex Syst.* **2,** 321 (1988).
32. J. Moody and C. Darken, "Fast learning in networks of locally tuned processing units," *Neural Computation* **1,** 281–294 (1989).

Harold H. Szu: Biography and photograph appear with the special section guest editorial.

Brian Telfer: Biography and photograph appear with the paper "Causal analytical wavelet transform" in this issue.

Shubha Kadambe received her undergraduate degrees in physics and electronics from Mysore University and Madras Institute of Technology, India, in 1977 and 1980, respectively. She received her MS (EE) from Tuskegee University, Alabama, in 1986 and her PhD (EE) from the University of Rhode Island in 1991. From 1980 to 1981, she was a trainee at Bhabha Atomic Research Center, a premier research organization in India. She was a scientific officer at the same organization from 1981 to 1984. She is currently a postdoctoral research fellow at the Applied Science and Engineering Labs, A. I. duPont Institute, Wilmington, Delaware, conducting research in developing speech aids for the handicapped. Her research interests include speech analysis and synthesis, speech modeling, speech enhancement, time-frequency and time-scale representations, neural networks, and image processing.