Speaker normalization of static and dynamic vowel spectral features

Stephen A. Zahorian and Amir J. Jagharghi
Department of Electrical and Computer Engineering, Old Dominion University, Norfolk, Virginia 23508-0369

(Received 16 October 1989; revised 6 August 1990; accepted 28 March 1991)

Two methods are described for speaker normalizing vowel spectral features: one is a multivariable linear transformation of the features and the other is a polynomial warping of the frequency scale. Both normalization algorithms minimize the mean-square error between the transformed data of each speaker and vowel target values obtained from a "typical speaker." These normalization techniques were evaluated both for formants and a form of cepstral coefficients (DCTCs) as spectral parameters, for both static and dynamic features, and with and without fundamental frequency ($F0$) as an additional feature. The normalizations were tested with a series of automatic classification experiments for vowels. For all conditions, automatic vowel classification rates increased for speaker-normalized data compared to rates obtained for nonnormalized parameters. Typical classification rates for vowel test data for nonnormalized and normalized features respectively are as follows: static formants—69%/79%; formant trajectories—76%/84%; static DCTCs 75%/84%; DCTC trajectories—84%/91%. The linear transformation methods increased the classification rates slightly more than the polynomial frequency warping. The addition of $F0$ improved the automatic recognition results for nonnormalized vowel spectral features as much as 5.8%. However, the addition of $F0$ to speaker-normalized spectral features resulted in much smaller increases in automatic recognition rates.

PACS numbers: 43.72.Fx, 43.70.Gr, 43.72.Ar

INTRODUCTION

The acoustic properties of phonologically equivalent vowels vary greatly because of coarticulation with adjacent phonemes and also because of individual speaker differences. Of these two sources of variations, the ones related to speaker effects are in general larger (Nearey, 1989). The speaker-dependent variations arise not only because of the physics of speech production but also because of systematic differences in pronunciation. For example, since children have shorter vocal tracts than adult males, the frequency components of children's speech are higher than are those of adult males. There are also consistent variations due to regional accents, speaking habits, etc. Despite the many apparent differences in acoustic cues, listeners can generally identify the vowel intended by the speaker. In contrast automatic methods for vowel classification have achieved more limited success in identifying vowels in a speaker-independent manner.

Automatic classification rates for vowels can be improved if the raw acoustic features are speaker normalized, using either intrinsic or extrinsic factors (Nearey, 1989), to account for the systematic differences among speakers. Intrinsic factors for normalization are sources of information contained within the vowel itself such as fundamental frequency ($F0$) or the third formant frequency ($F3$) or some combination of the two. Several investigators have utilized intrinsic information to improve automatic vowel classification accuracy [for example, Wakita (1977); Syrdal and Gopal (1986); Miller (1989); Hillenbrand and Gayvert (1987)]. In other studies, vowel normalization has been attempted with extrinsic factors such as a speaker-dependent frequency scaling or the incorporation of additional features that are dependent on the global characteristics of each speaker [Gerstman (1968), Nearey (1978), Hindle (1978)]. Nearey (1989) concluded that human listeners themselves use both intrinsic and extrinsic factors for vowel normalization. Disner (1980) evaluated several vowel normalization techniques both in terms of speaker-related variance reduction in $F1/F2$ vowel plots and also in terms of their effects in preserving significant linguistic differences among vowel tokens.

In this paper, we present two algorithms for extrinsic speaker normalization of vowel spectral features and give experimental results for the two techniques. One normalization is a linear transformation of spectral features, i.e., a matrix multiplication, with a separate transformation computed for each speaker. The other normalization is a speaker-dependent frequency warping. In both cases, the normalization parameters minimize the mean-square error between the normalized spectral features for each vowel and the "target" position for that vowel. These normalization methods were tested using automatic classification experiments using two feature sets: formants and DCTCs, a form of cepstral coefficients. Automatic classification results based on speaker-normalized spectral features were compared with results obtained with nonnormalized features, nonnormalized features plus $F0$, and a combination of normalized features plus $F0$. Each extrinsic normalization technique was evaluated as a function of the number of vowels used to compute the normalization for both static and dynamic spectral features. The normalization techniques presented in this paper are an extension of methods previously presented in the literature.

I. DATABASE

The database for the experiments was obtained by recording 99 CVC syllables produced in isolation by each of 30 speakers. Appendix A lists these CVCs. Ten of the speakers were adult males (M), 10 were adult females (F), and 10 were children (C) between the ages of 7 and 11 (5 male, 5 female). These speakers were all native speakers of English, but had no formal phonetic training. Approximately half of the speakers were natives of Virginia; the other half were natives of other geographical regions of the United States, predominantly the Northeast. The CVC syllable list contained approximately 9 instances of each of the 11 vowels /a,i,u,æ,ɝ,ɪ,ɛ,ɔ,ʌ,ʊ,o/. The initial consonant was one of /b,d,g,p,t,k,h,l,w/. The final consonant was one of /b,d,g,p,t,k,v,s/.

Since the speakers were phonetically naive, the nonword CVCs were reviewed with each speaker as to the intended pronunciation. In the recording session, each syllable was displayed on a computer monitor using an English-letter pseudophonetic spelling, as given in Appendix A. The listeners repeated syllables that the experimenter[1] judged to be mispronounced. The typical sound level of speech sounds was approximately 36 dB above the background noise level in the sound proof room used for the recordings. The speech signals were lowpass filtered at 7.5 kHz and sampled at 16 kHz with a 12-bit analog-to-digital converter. All recordings were then digitally filtered at 240 Hz with a 62nd-order FIR linear phase high-pass filter to remove low-level low-frequency noise in the signal. For each token, the experimenters manually labeled the transition to the vowel, the steady-state vowel region, and the final transition. They also deleted stimuli from the database that they judged to be mispronounced. The total number of accepted vowel stimuli was 2922 (out of $99 \times 30 = 2970$).

II. SPEECH PARAMETERS AND CLASSIFICATION METHODS

A. Speech parameters

The normalization techniques discussed in this paper were investigated with two raw parameter sets—formants and a form of cepstral coefficients. These two feature sets enable a more comprehensive evaluation of normalization techniques than would a single feature set. The algorithms presented in this paper are very general in that they could also be used with any other feature set derived from vowel spectra.

Formants were computed in a multistage process as follows. The speech signal was first digitally low-pass filtered at 3.8 kHz with a 49th-order FIR linear-phase low-pass filter and resampled at 8 kHz. The speech signal was then high-frequency preemphasized with transfer function $1 - 0.75\ z^{-1}$. The signal was windowed with a 25-ms Hanning window and a 10th-order LP model was computed. The roots of the LP polynomial were computed to obtain up to five formant candidates per frame. Formant candidates were computed for nine frames of the vowel, with frames equally spaced throughout the steady-state portion of the vowel. Finally a formant tracking routine (similar to McCandless, 1974) was used to track the first three formants over the duration of the vowel. The resulting formant values for the fifth frame were selected as the three formant values for each vowel.

The cepstral coefficients were computed as the d_j in the equation:

$$H'(f') = \sum_{j=1}^{j=N} d_j \cos[(j-1)\pi f']. \qquad (1)$$

The H' implies nonlinear amplitude scaling of the magnitude spectra, and f' implies a nonlinear frequency scaling. The frequency scale was normalized so that a selected frequency range in f of $f_1 \leqslant f \leqslant f_2$ corresponds to $0 \leqslant f' \leqslant 1$. In our experiments the d_j's were computed after first high-frequency preemphasizing the signal $(1 - 0.95\ z^{-1})$, using a 25-ms Hamming window, and computing the magnitude spectra for each frame. Several experiments were conducted to evaluate various nonlinear amplitude scales, nonlinear frequency scales and frequency ranges, in terms of their effect on automatic vowel recognition accuracy without speaker normalization. Based on these experiments, a log amplitude scaling was used and bilinear frequency warping (Oppenheim and Johnson, 1972) with a coefficient of 0.6 was used. That is,

$$f' = f + \frac{1}{\pi} \tan^{-1}\left\{\frac{0.6 \sin(2\pi f)}{1 - 0.6 \cos(2\pi f)}\right\}. \qquad (2)$$

The cepstral coefficients were computed from the original speech signal (i.e., no decimation and low-pass filtering), but over a frequency range of 150–5000 Hz. Automatic classification experiments without speaker normalization also indicated that highest automatic recognition results for test data were obtained using cepstral coefficients 2–9. Since the cepstral coefficients were computed as a discrete cosine transform of a selected segment of the nonlinearly scaled magnitude spectrum, we refer to them as DCT coefficients or DCTCs throughout the remainder of this paper.

Fundamental frequency ($F0$) was computed using a form of the SIFT fundamental frequency algorithm (Markel, 1972). That is, the LP residual was computed for a window of speech (50 ms for males, 40 ms for females and children) in the steady-state portion of each vowel with a 12th-order LP inverse filter. Here, $F0$ values were computed from peaks in the autocorrelation of the residual after low-pass filtering at 1 kHz. The details of the signal processing for the $F0$ extraction, including the LP window lengths, were developed and investigated in previous studies (Zahorian and Gordy, 1983; Effer, 1985). For dynamic features, $F0$ values were computed for each frame of the vowel.

B. Dynamic features

Speech features were also computed for each of several speech frames, in order to evaluate automatic recognition accuracy of both nonnormalized and speaker-normalized spectra for the case of dynamic spectra. Based on pilot experiments, the approach used for the data given in this paper was to first sample the speech spectra with 25 frames equally spaced from the beginning of the initial transition through the end of the final transition. Thus the frame spacing depended on the stimulus duration. The values of each parameter over the 25 frames (i.e., a vector of length 25) was then expanded in a three-term cosine basis vector expansion. The first basis vector was a constant, the second basis vector one-half cycle of a cosine, and the third basis vector was one cycle of a cosine. The coefficients of this cosine expansion were then used as the input to the classifier. Thus dynamic features consist of time-smoothed and time-normalized parameter trajectories. The total number of dynamic features is 3 × (number of parameters per frame), or 9 for the case of formants and 24 for the case of the DCTCs. A similar procedure has been used in studies with consonants (Tanaka, 1981).

C. Classifier

All speaker-normalization methods were evaluated in terms of their effects on automatic vowel classification with a Bayesian maximum likelihood (MXL) classifier. That is, each stimulus was classified as a member of the category for which the distance,

$$D_i(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_i)^T \mathbb{R}_i^{-1} (\mathbf{x} - \mathbf{x}_i) + \ln|\mathbb{R}_i| - 2\ln \mathbb{P}(G_i), \quad 1 \leqslant i \leqslant G, \tag{3}$$

is minimized. In Eq. (3), $\mathbf{x}_i$ is the centroid for category G_i, $\mathbb{R}_i$ is the covariance matrix for category G_i, $\mathbb{P}(G_i)$ is the *a priori* probability for category G_i, and G is the number of categories (i.e., vowels). Thus each category is characterized according to the centroid of all the data in that category and the covariance matrix of the data in that category. This classifier is optimum if the feature vector components are multivariate Gaussian (Duda and Hart, 1973).

III. NORMALIZATION PROCEDURES

A. Linear transformation method

In general terms, a linear transformation was computed for each speaker such that transformed features for each speaker were as similar as possible to the feature vectors for a single "typical" speaker. This type of normalization can be applied to any feature set computed from the vowel signal including formants and DCTCs. In mathematical terms, the method consists of a speaker-specific linear transformation $\mathbb{T}$ and offset $\mathbf{o}$ from raw feature vectors $\mathbf{x}$ to normalized feature vectors $\mathbf{y}$, i.e.,

$$\mathbf{y} = \mathbb{T}\cdot\mathbf{x} + \mathbf{o}. \tag{4}$$

Thus each component of the speaker-normalized feature vector $\mathbf{y}$ is a linear combination of the original feature vector components plus an offset term.

The matrix $\mathbb{T}$ and vector $\mathbf{o}$ are computed such that normalized feature vectors are as similar as possible, in a mean-square error sense, to the target position for the corresponding vowel. That is, the elements in $\mathbb{T}$ and $\mathbf{o}$ for each speaker minimize

$$E = \sum_{i=1}^{G} \sum_{k=1}^{NS(i)} \sum_{j=1}^{N} (\mathbf{x}'_{ji} - \mathbf{y}_{jik})^2, \tag{5}$$

where $\mathbf{x}'_{ji}$ = jth component of the "target" position for the ith vowel, $NS(i)$ = number of training stimuli for vowel i, and $\mathbf{y}_{jik}$ = kth stimulus, component j for vowel i. Target positions must first be computed for each vowel, as discussed in more detail in a latter section. In Eq. (5), we assume there are G vowels and N components in each feature vector. Mean-square estimation methods can be used to compute the elements in $\mathbb{T}$ and $\mathbf{o}$ so as to minimize E in Eq. (5) (Zahorian and Jagharghi, 1991).

It should be noted that this linear transformation algorithm is quite different from a linear scaling of the frequency axis (with frequency measured in either Hz, log Hz, or Barks), as used in some previous studies as a normalization technique. For example, the linear normalization presented by Golibersuch (1983) is a linear frequency scaling. In contrast, the technique presented in this paper is a general linear transformation or matrix multiplication of the original vowel features. The large number of variables in the normalization matrix, in contrast to a single variable for a linear frequency scaling, provides many more degrees of freedom for characterizing the detailed behavior of each speaker. However, as with any statistically based parameter estimation procedure, a model with a large number of variables might not generalize as well to data outside the training set as would a model with a small number of parameters. Our objectives in this study were to determine not only a transformation with good performance, in terms of vowel recognition rates for test data, but also to determine a transformation that could be computed with a minimum of training data.

In order to assess the linear normalization described, the technique was tested with both formants and DCTCs as the feature vector components. The normalization was also evaluated both with and without the offset term in Eq. (4). The number of vowels used to compute the normalization coefficients was varied from 3 to 11 (see footnote 2). The normalization coefficients were also computed for cases such that only terms on the diagonal in $\mathbb{T}$, or the diagonal in $\mathbb{T}$ plus elements adjacent to the diagonal on each side, were allowed to be nonzero. For example, for the case of the diagonal only, each normalized feature is simply a scaled version of the corresponding unnormalized feature. Each normalization condition was evaluated with the MXL classifier mentioned, applied to all 11 vowels.

B. Frequency-warping method for speaker normalization

Another method of speaker normalization incorporates a speaker-dependent frequency-warping function for each speaker. Unlike the linear transformations described above, this normalization is motivated from a physical basis—that is, speaker-dependent differences in vocal tract lengths cause a scaling in the frequency domain. For our experiments, frequency warping was only used with DCTCs as parameters. For this algorithm, speaker-dependent coefficients a, b, and c are to be determined so as to redefine the already-warped frequency scale for each speaker according to

$$f'' = af'^2 + bf' + c. \tag{7}$$

The normalized DCTCs are the d_j's in the equation,

$$H'(f'') = \sum_{j=1}^{N} d_j \cos[(j-1)\pi f''], \quad \text{for } 0 \leqslant f'' \leqslant 1 \qquad (8)$$

The a, b, and c values are chosen for each speaker to minimize

$$E = \sum_{i=1}^{G} \sum_{k=1}^{NS(i)} \sum_{j=2}^{N} (d'_{ji} - d_{jik})^2, \qquad (9)$$

where d'_{ji} is the target value of d_j for vowel i and d_{jik} is the d_j for the kth training stimulus of vowel i.

Thus the speaker-normalized DCTCs are computed with a speaker-normalized frequency scale. Note that the speaker-dependent warping occurs after the fixed bilinear frequency warping mentioned previously [Eq. (2)]. The second-order polynomial allows more flexibility in the warping function than would be possible through a linear scaling of the frequency axis, while parametrizing the warping function with only three coefficients. This method was used with DCT coefficients 2–9 as the feature vector. For each speaker the a, b, and c values were then determined such that DCT coefficients computed with the warped frequency scale would match the target positions as closely as possible. A speaker-dependent frequency range, selected from a range of 0–8000 Hz for each speaker, was thus mapped to a range of 150–5000 Hz (used for the target speaker) with the polynomial mapping given in Eq. (7). A gradient search was used to solve the nonlinear optimization problem for the coefficients a, b, and c for each speaker. This method is somewhat similar to frequency-warping methods previously reported in the literature. Two of these previous studies (Paliwal and Ainsworth, 1985; Neuburg 1988) used frequency warping as intrinsic normalization in that the warping function was computed as a by-product of distance calculations between unknown and reference spectra. In another study (Neuburg, 1980) both linear and nonparametric nonlinear frequency scalings were used to match formant histograms derived from a sample of each speaker's speech. Although these implementations of frequency warping did improve spectral matching, no evidence was given for improved automatic recognition rates for either vowels or isolated words.

C. Normalization based on $F0$

Normalization based on $F0$ was straightforward—$F0$ was simply included as an additional parameter for the classifier. Although some researchers have modified their original parameters according to $F0$ values [for example, Syrdal and Gopal (1986); Miller(1989)], other researchers have obtained improved automatic recognition results simply by augmenting the feature vector with the $F0$ value [Hillenbrand and Gayvert (1987); Syrdal (1985)]. Here, $F0$ normalization was tested both with and without the two types of extrinsic normalization.

D. Selection of target positions for each vowel

Target positions for each vowel were computed as follows. The first "estimate" of the target for each vowel is simply the average of all the nonnormalized training data for that vowel. The averages were computed using

$$x'_{ji} = \frac{1}{NS(i)} \sum_{k=1}^{NS(i)} x_{jik}, \quad 1 \leqslant i \leqslant G, \quad 1 \leqslant j \leqslant N, \qquad (6)$$

where $NS(i)$, G, and N are as defined in a previous section. Each speaker was then evaluated in terms of the mean-square error match of that speaker's training data to the average-value targets. The speaker with the best match to the overall speaker averages was considered to be the most "typical" speaker. The final targets were the category averages for this most typical speaker. Based on automatic vowel recognition performance in pilot experiments, we chose the averages from the typical speaker versus the targets from Eq. (6) as the targets for the experiments reported in this paper.

IV. EXPERIMENTS

For all automatic vowel recognition experiments, approximately half the database were used to train the classifier and the other half were used to test the classifier. For increased statistical reliability, the training and test data were then interchanged and results are given as the average of the two sets of test results. Only test results are given since these results represent the ability of the classifier to generalize, unlike the training results. For the case of speaker normalization, approximately half the normalized data of *each speaker* were used for training the classifier and the other half for testing the classifier. A portion of each speaker's training data, varying from 3 to 11 vowels, were used to determine a normalization transformation for that speaker.[2] The vowels used for computing the normalization were ordered as given in the list of 11 vowels in a previous section, and as repeated in Table I.[3] Test classification results thus were derived from data that is independent of both data used to train the classifier and to compute the speaker normalizations. As a result of pilot tests, the normalization conditions listed in Table II were selected for the experimental evaluations reported in this paper. The table also lists a code for each condition for use in subsequent tables, figures, and discussion.

A. Steady-state vowel formants

1. Control results

Table III depicts training and test automatic recognition results for four control cases for formants. For case A, the training and testing speakers are different. For this case five adult females, five adult males, and five children were

TABLE I. Order of vowel usage for computing normalizing coefficients.

IPA	DARPABET	Example
a	aa	cot
i	iy	beat
u	uw	boot
æ	æ	bat
ɝ	er	bird
ɪ	ih	bit
ɛ	eh	bet
ɔ	ao	bought
ʌ	ah	bud
ʊ	uh	book
o	ow	home

TABLE II. Explanation of codes used in paper.

NN	No normalization.
NN + $F0$	No normalization, $F0$ included as an extra feature.
PN	Polynomial frequency warping normalization.
PN + $F0$	Polynomial frequency warping normalization, plus $F0$.
PX	Partial matrix linear normalization. For the case of formants, a diagonal matrix and no offset term. For the case of DCTCs, a matrix with the main diagonal plus 1 term on each side of the diagonal plus an offset vector.
PX + $F0$	Same as PX, plus $F0$.
FX	Full matrix linear normalization plus an offset term.
FX + $F0$	Same as FX, plus $F0$.

used for training with the remaining speakers used for testing. For case B, half the data of each speaker were used for training the classifier and the other half for testing. Thus this case uses the same data management for the classifier as was required for the speaker normalization cases, but does not use explicit speaker normalization. Case C is the same as case A, i.e., independent training and test speakers, except $F0$ is included as an additional parameter. Case D is the same as case B, except for the addition of $F0$ as a parameter. For cases A and C results are also given for the three speaker types in the study. The results for the males are slightly higher than for the females which in turn are slightly higher than for the children; the difference between the males and children is 2.3% for case A and 5.6% for case C. The results averaged over all speakers range from 66.9% to 75.1% percent correct depending on the condition. It thus can be seen that speaker type, data management, and the addition of $F0$ as a parameter affect the classification rates.

2. Normalization results

Table IV lists test classification results for speaker-normalized formants as a function of the number of vowels used to determine the speaker-normalization transformations for both the PX and FX conditions. For PX, the normalization transformations are diagonal matrices with no offset. Thus each speaker-normalized formant is simply a scaled version of the nonnormalized formant. For FX, each normalized formant is a linear combination of all three unnormalized formants plus an offset term. Thus 3 coefficients must be computed per speaker for the PX case versus 12 for the FX case. Normalization based on 0 vowels, which is case B from Table III, i.e., no normalization at all, is listed in Table IV for comparison. On the average the PX results are slightly better than the FX results.[4] The PX results improve even if only the three vowels /a,i,u/ determine the normalization coefficients for each speaker. Note that a small additional improvement results if $F0$ is included as an extra parameter along with the normalized formants. Since normalization based on speaker-dependent formant scaling (PX) gives comparable results to the full matrix case (FX), and since the small number of PX coefficients can be computed with less data than for the FX case, the PX normalization appears to be preferable to the FX normalization.

TABLE III. Automatic vowel classification rates for 11 vowels based on three formants without explicit speaker normalization.

Case	Separate speakers (Trn vs Tst)	$F0$?	Males	Females	Children	All
A	Yes	No	68.3%	66.6%	66.0%	66.9%
B	No	No	···	···	···	69.3%
C	Yes	Yes	74.1%	73.5%	68.5%	72.1%
D	No	Yes	···	···	···	75.1%

TABLE IV. Automatic classification rates for 11 vowels based on three speaker-normalized formants.

No. of normalizing vowels	PX	FX
0	69.3%	69.3%
3	74.5%	[a]
5	76.3%	75.2%
7	77.2%	74.9%
9	78.0%	79.2%
11	78.6%	79.0%
11 + $F0$	80.5%	80.4%

[a] See footnote 2.

The $F1/F2$ plane cluster plots depicted in Fig. 1 also

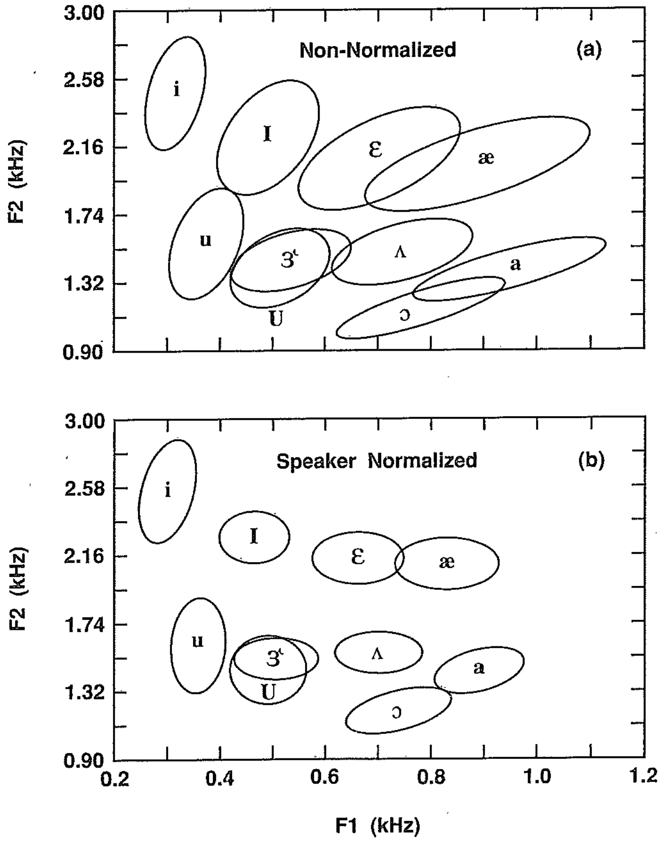

FIG. 1. Cluster plots of ten vowels for (a) unnormalized formant data and (b) speaker-normalized formant data.

illustrate the effects of PX speaker normalization for the vowel formant data. For each vowel, an ellipse was computed with its major axis oriented in the direction of maximum data variation and equal to two standard deviations of the data in this direction. Similarly the minor axis equals two standard deviations of the data in the direction orthogonal to the major axis. For Fig. 1(a), the ellipses represent the original formant values. For Fig. 1(b), the ellipses depict data for the speaker-normalized formants. The vowels are clearly better clustered in panel b than panel a. Note that for both plots /o/ was not used since its $F1/F2$ values overlap greatly with /u/ and /ɝ/ both before and after speaker normalization. For Fig. 1(b), 50% of each speaker's data was used to compute the speaker normalization. This normalization was then applied to all the data of each speaker.

Numerical figures of merit were also computed to evaluate the data shown in Fig. 1. First the within-class variance for Fig. 1(b) is 46% of the within-class variance for Fig. 1(a). Secondly, as per a method given in Syrdal (1985), a classification experiment was performed to determine the accuracy with which speaker groups (M/F/C) could be identified from the formant data before and after normalization. For the non-normalized formants ($F1$ and $F2$), the speaker group was correctly identified for 57.2% of data. For the normalized data, the speaker group was identified only for 39.8% of the data. In a similar experiment with all three formants, speaker groups could be identified with 73.0% and 40.1% accuracies for the non-normalized and normalized formants respectively. Thus speaker types can be identified from the vowel data at substantially above chance (33.3%) only with the original formants.

B. Normalization for DCTC's computed from the steady-state vowel

1. Control

Table V lists control results for the same conditions as used for Table III, except eight DCT coefficients (2 to 9) encoded each vowel stimulus. Once again the test results depend on speaker type, on whether $F0$ is included as a parameter, and on whether testing data are from the same or different speakers as were used for the training data. For the all-speaker case, classification rates range from 69.6% to 80.2%. All the rates, except for female speakers in case C, are somewhat higher than corresponding rates given in Table III for the formants.

TABLE V. Automatic vowel classification rates for 11 vowels based on 8 DCTCs without explicit speaker normalization.

Case	Speakers separate (Trn vs Tst)	$F0$?	Males	Females	Children	All
A	Yes	No	68.6%	71.2%	69.0%	69.6%
B	No	No	⋯	⋯	⋯	74.8%
C	Yes	Yes	74.5%	72.2%	71.1%	72.6%
D	No	Yes	⋯	⋯	⋯	80.2%

2. Normalization results

Table VI lists test classification results for speaker-normalized data as a function of the number of vowels used to obtain the normalization coefficients for each speaker. Results are given for three normalization conditions: PN, polynomial normalization; PX, a linear transformation consisting of a matrix with diagonal elements and elements on either side of the diagonal, plus an offset; and FX, a linear transformation consisting of an 8×8 matrix plus an offset. The number of coefficients required to normalize the data of each speaker is 3 for PN, 30 for PX, and 72 for FX. The recognition rates are somewhat higher for PX and FX versus PN. However, more vowel data is required to maximize performance for the PX and FX cases versus the PN normalization. In fact, if seven or fewer vowels are used to compute the normalization, the best performance is obtained with PN. The addition of $F0$ improves performance somewhat for PN and PX, but very little for the FX case. For the best case (PX), approximately 84% of test vowels were correctly classified without the use of $F0$ versus 85% with $F0$ included.

To illustrate the effects of the linear-transformation normalization on speech spectra, several spectral plots were made of vowel spectra before and after speaker normalization. These plots depict vowel data outside the set used to compute the speaker normalization. Figure 2 depicts the spectrum for the vowel /a/ before and after FX normalization as well as the target spectrum used for this vowel. Similarly, Fig. 3 depicts the result of the PN normalization. The figures show that the speaker-normalized spectra is much more similar to the target spectrum than are the original spectra. The effect is more pronounced for the FX normalization than for the PN normalization. Other examples yield similar results—thus illustrating the effectiveness and generality of the transformations.

Figure 4 depicts the average frequency warping curves obtained for adult males, adult females, and children. For the adult male speakers, the frequency range of 84–4115 Hz is mapped to the target frequency range of 150–5000 Hz. Since the "typical" speaker for computing target vowel positions was a female, on the average, normalization for the female speakers has very little effect. For the children, the frequency range of 165- to 6735-Hz maps to the 150- to 5000-Hz range. These values are consistent with expected frequency differences for these speaker types.

TABLE VI. Automatic classification rates for 11 vowels based on 8 speaker-normalized DCTCs.

No. of normalizing vowels	PN	PX	FX
0	74.8%	74.8%	74.8%
3	77.1%	[a]	[a]
5	78.9%	76.2%	[a]
7	79.6%	78.4%	[a]
9	78.8%	82.8%	81.1%
11	78.7%	84.2%	83.8%
$11+F0$	82.2%	85.2%	84.1%

[a] See footnote 2.

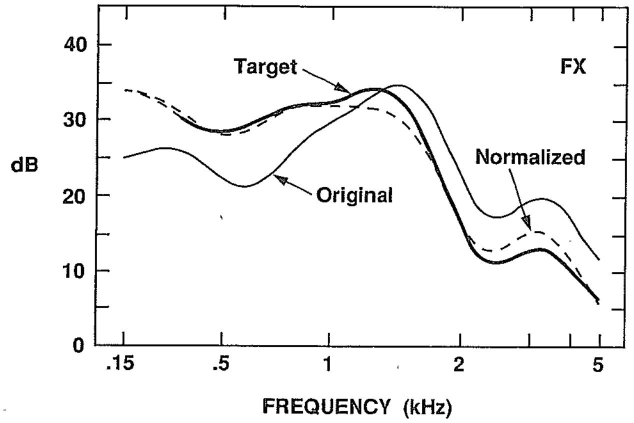

FIG. 2. Spectral plots for the vowel /a/ for a "typical" speaker, the original spectrum of a child's /a/, and the speaker-normalized spectrum. The speaker normalization was computed as a linear transform of eight DCT coefficients.

C. Speaker normalization of dynamic spectra

With both formants and DCTCs as original parameters, speaker normalization of several frames of spectral parameters was investigated in an attempt to obtain yet additional improvements in automatic vowel recognition. In all cases, the normalization coefficients were computed from a single static frame, located in the steady-state portion of the vowel. All static training data for all 11 vowels were used to compute the normalization. The normalization transformation was applied to all 25 frames of parameters, computed as mentioned previously. Each normalized parameter was then encoded as the coefficients in a 3-term cosine basis vector expansion over the 25 frames for that parameter.

Typical automatic classification results for formants and DCTCs are given in bargraph form in Fig. 5 and 6, respectively. For comparison, representative results for static spectra are also given in Figs. 5 and 6, again with normalizations based on all 11 vowels. The NN (no normalization) results for the static and dynamic cases show that dynamic information increases the test recognition rates for both formants and DCTCs even without speaker normalization.[5] Speaker normalization of the dynamic features (conditions PN, PX, FX), or the addition of $F0$ (NN + $F0$), improves the recognition rate even more. Note that only the average (over time) of $F0$ was used for classification, since pilot tests indicated that the trajectory of $F0$ did not improve classification. The effect of $F0$ is comparable to PN normalization, but less effective than the linear transformation normalizations (PX and FX). However, the addition of fundamental frequency to speaker-normalized parameters improves recognition for the polynomial frequency warping (PN + $F0$) but improves recognition only slightly for the linear transformation normalization based on a partial matrix (PX + $F0$) and even degrades recognition slightly for the full matrix linear transformation (FX + $F0$). The best test rate for DCTC parameters is 90.8% (PX + $F0$) whereas the best test rate based on formants is 84.2% (FX + $F0$).

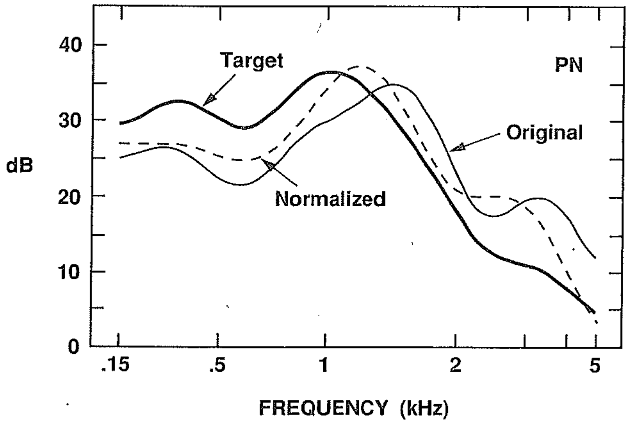

FIG. 3. Spectral plots for the vowel /a/ for a "typical" speaker, the original spectrum of a child's /a/, and the speaker-normalized spectrum. The speaker normalization was computed using polynomial frequency warping.

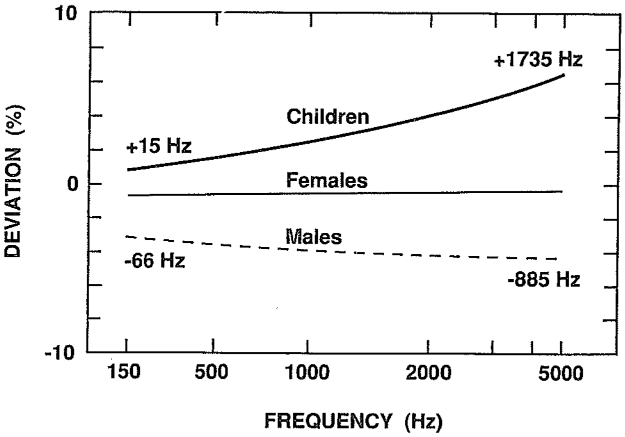

FIG. 4. The speaker-dependent frequency-warping functions obtained for speaker normalization, as averages over speaker categories. The original frequencies were computed with respect to a bilinear frequency-warping function. Results are shown in terms of deviations from the original frequencies. The frequency scale for each speaker was warped to best match the "typical" speaker over a range of 150–5000 Hz.

V. DISCUSSION AND CONCLUSIONS

Speaker-normalization techniques have been described to reduce systematic speaker differences in acoustic features for vowels. The primary objective of this paper was to present a mathematically based class of normalization techniques and experimentally demonstrate these algorithms with two feature sets, rather than to compare the two feature sets. These normalization techniques were: FX, a linear transformation consisting of a full transformation matrix; PX, a linear transformation consisting of a diagonal "strip" in the transformation matrix; and, PN, a polynomial warping of the frequency axis. Note that PX is in effect a special case of FX. These normalization techniques could also be applied to other feature sets for vowels. In this sense, the algorithms presented in this paper are more general than most speaker-normalization techniques previously reported in the literature.

The normalization algorithms were experimentally evaluated under a large number of conditions for two parameter sets—modified cepstral coefficients (DCTCs), which encode overall spectral shape, and formants, which encode the spectral peaks.[6] In all cases slightly higher automatic

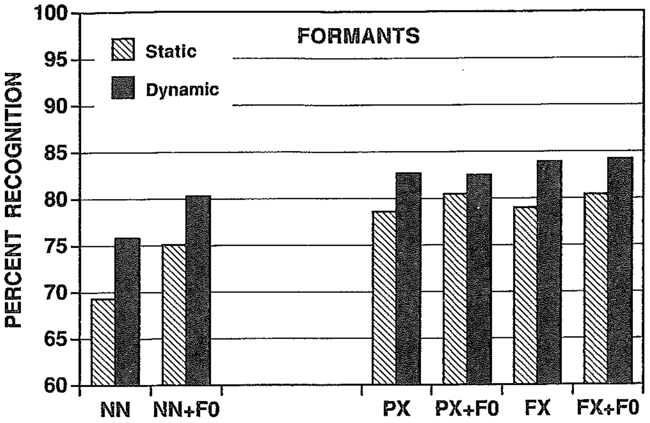

FIG. 5. Automatic recognition rates for 11 vowels based on formants for different processing methods as noted.

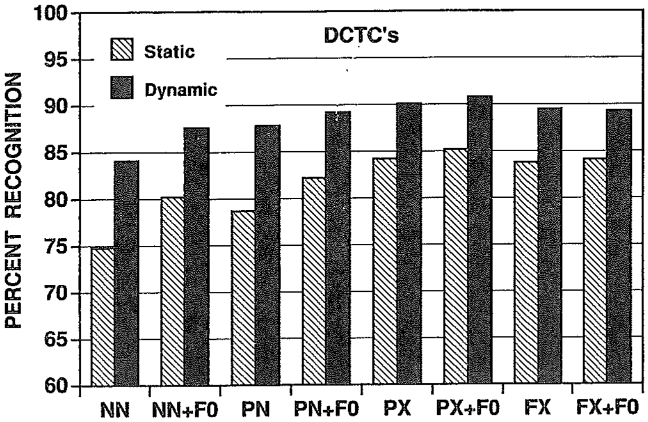

FIG. 6. Automatic recognition rates for 11 vowels based on DCTCs for different processing methods as noted.

classification rates were obtained with the DCTCs than with formants. However, the DCTCs spanned a greater frequency range than did the three formants, 150–5000 Hz versus 150–3800 Hz and comprise a much higher dimensionality feature space than do the formants (eight coefficients versus three formants). There is also, of course, the possibility and likelihood of at least some errors in automatic formant tracking. However, under the assumption that the formants were tracked correctly for adult males (generally the easiest case), and since the magnitude of the differences in formant vowel classification rates for children versus adult males was generally less than the magnitude of differences between formant and DCTC results, it seems unlikely that all differences in DCTC versus formant results can be attributed to tracking errors.[7]

For formant vowel normalization, the PX and FX linear transforms generally resulted in very similar classification rates. However, the PX transform required fewer parameters per speaker than the FX transform (3 vs 12). Although the addition of $F0$ to unnormalized formants resulted in improved classification results, the addition of $F0$ to normalized formants resulted in only very small improvements, thus indicating that the $F0$ information is correlated with the speaker-normalization coefficients.

Of the speaker-normalization techniques investigated for use with the DCTC parameters, the PX linear transform method was generally superior to both the FX linear transform and the PN polynomial frequency warping. Apparently the PN method, with only three coefficients per speaker, was not adequate to fully eliminate speaker differences. The FX transform apparently had too many coefficients (72 vs 30 for PX) for adequate trainability.[8] The PN + $F0$ normalization was almost as effective in increasing automatic vowel recognition rates as were the linear transform methods. However, as for the formants, the addition of $F0$ to features normalized with the linear transform methods did not substantially improve automatic recognition rates. Automatic classification rates obtained with speaker-normalized vowel data as represented by dynamic spectral shape are very similar to identification rates obtained by human listeners.

Because of differences in data bases and methodology, only general comparisons can be made between the normalization algorithms presented in this paper and those previously reported in the literature. The variance reduction for the PX formant normalization exceeds all those listed in Disner (1980) for English vowels, except for the log-mean (Nearey, 1978) normalization. The normalization algorithms presented in this paper are more successful in improving automatic vowel classification than are previously reported normalization techniques. Using the experimental data that is most directly comparable to our tests, Wakita (1977) improved vowel recognition from 78.9% to 84.4% for unnormalized versus normalized formants, a smaller percentage improvement than obtained with our PX normalization of formants (69.3% to 78.6%). The frequency warping speaker-normalization method reported by Paliwal and Ainsworth (1985) degraded automatic vowel and word classification. However, unlike the intrinsic normalization method in both the Wakita and the Paliwal and Ainsworth study, our extrinsic method requires not only that speaker identity be known, but that known vowel samples be available from each speaker in order to compute the normalization transformation for that speaker. We did show, however, that knowledge of only three to five of each speaker's vowels is sufficient to compensate for the primary speaker-dependent effects.

Clearly more sophisticated techniques could be devised to account for systematic speaker variations in dynamic vowel spectral features. The normalizations given in this paper for the dynamic spectra are based solely on the static spectrum and do not consider the possibility of systematic timing differences between speakers. Nevertheless, classification results derived from normalized dynamic spectra are very similar to human perception of the vowel stimuli (Jagharghi, 1990), thus indicating the procedures given appear to account for the primary speaker differences in dynamic vowel spectra.

ACKNOWLEDGMENT

This work was supported by Grants IRI-8702649 and IRI-9003226 from the National Science Foundation.

APPENDIX A

List of 99 CVC syllables recorded for vowel database. The DARPABET phonetic spelling for each syllable is also given.

pot/paat/
got/gaat/
pock/paak/
teak/tiyk/
peeb/piyb/
boot/buwt/
coop/kuwp/
moog/muwg/
tack/taek/
tab/taeb/
dirt/dert/
gerp/gerp/
tick/tihk/
bid/bihb/
bet/beht/
get/geht/
bought/baot/
talk/taok/
gawp/gaop/
putt/paht/
cub/kahb/
took/tuhk/
hood/hhuhd/
code/kowd/
goag/gowg/
bah/baa/
top/taap/
hod/hhaad/
deep/diyp/
keyed/kiyd/
poop/puwp/
gook/guwk/
who'd/hhuwd/
dad/daed/
tag/taeg/
curb/kerb/
durg/derg/
kit/kiht/
pig/pihg/
ted/tehd/
web/wehb/
caught/kaot/
paup/paop/
hawd/hhaod/
dug/dahg/
bud/bahd/
put/puht/
boat/bowt/
pope/powp/
coab/kowb/
tog/taag/
cob/kaab/
peep/piyp/
keep/kiyp/
league/liyg/
toot/tuwt/
tube/tuwb/
pat/paet/
cap/kaep/
had/haed/
perk/perk/
heard/hherd/
give/gihv/
hid/hihd/
debt/deht/
peck/pehk/
daub/daob/
baud/baod/
but/baht/
cup/kahp/
hud/hhahd/
could/kuhd/
dope/dowp/
toad/towd/
hoed/hhowd/
dot/daat/
pod/paat/
beet/biyt/
geese/giys/
heed/hhiyd/
dupe/duwp/
sued/suwd/
bat/baet/
gap/gaep/
bird/berd/
turk/terk/
dip/dihp/
bib/bihb/
pep/pehp/
keg/kehg/
head/hhehd/
gawk/gaok/
cawg/kaog/
tuck/taht/
gut/gaht/
book/buhk/
good/guhd/
goad/gowd/
coke/kowk/

[1] The experimenters were the authors of this paper. One is a native US citizen (S. A. Zahorian) and the other has been in the US over 12 years (A. J. Jagharghi). Both have several years of previous experience with vowel experiments.

[2] However, the number of vowels used to compute the normalization coefficients must be at least one more than the number of coefficients in each row of the normalization matrix to insure that the speaker-normalized features are linearly independent, a requirement for computation of the matrices for the MXL classifier [Zahorian and Jagharghi (1990)].

[3] This order was chosen so that the more widely separated vowels, e.g., /a,i,u/ would appear first in the list. We assumed that knowledge of the range of a speaker's vowel space would enable a better estimate of normalization parameters than would knowledge of a small portion of that speaker's vowel space. However, we did not test this assumption using other vowel orders.

[4] Tests of statistical significance were not performed to compare every set of results obtained in this study. However, an estimate of the statistical significance of the results was obtained as follows. For the static NN conditions for formants and DCTCs, a one-tailed t test implied that differences in classification results of 2.1% are statistically significant at the 95% level of confidence. Assuming that normalization reduces speaker-related variance by approximately 50% (as for the data in Fig. 1), differences in classification results of 1.5% are statistically significant at the 95% level of confidence for the normalization conditions (or 2.1% at the 99% confidence level).

[5] Inspection of Figs. 5 and 6 shows that the difference between static and dynamic vowel classification rates, with all other conditions constant, is generally between 5% and 10%. In pilot experiments (condition NN for formants and DCTCs), we found that if the diphthong /o/ is removed from the vowel data, the differences between the static and dynamic features were only 1.2% and 4.3% for formants and DCTCs, respectively (versus 6.1% and 9.3% for these two cases with /o/ included). Thus presumably most, but not all, of the difference between static and dynamic feature performance is related to /o/.

[6] The PN method was not used for formants since it would have been essentially the same as the PX method for these features.

[7] In contrast, however, there may have been significant $F0$ errors for the children. Inspection of Tables III and V shows that the addition of $F0$ improves control vowel recognition much less for children than for adult males for both formants and DCTC features.

[8] For both formants and DCTCs, the FX transform with maximum flexibility gave the highest recognition results for training data. However test recognition results degrade somewhat compared with the PX results, suggesting that the FX transformations were overly tuned to the training data.

Disner, S. F. (**1980**). "Evaluation of vowel normalization procedures," J. Acoust. Soc. Am. **67**, 253–261.

Duda, R. O., and Hart, P. E. (**1973**). *Pattern Analysis and Scene Classification* (Wiley, New York, 1973).

Effer, E. A. (**1985**). "An investigation to improve linear predictive vocoder pulse excitation models," Master's thesis, Old Dominion University.

Gerstman, L. (**1968**). "Classification of self-normalized vowels," IEEE Trans. Audio Electro. Acoust. **16**, 78–80.

Golibersuch, R. J. (**1983**). "Automatic prediction of linear frequency warp for speech recognition," ICASSP **83**, 769–772.

Hillenbrand, J., and Gayvert, R. T. (**1987**). "Speaker-independent vowel classification based on fundamental frequency and formant frequency," J. Acoust. Soc. Am. Suppl. 1 **81**, S93.

Hindle, D. (**1978**). "Approaches to vowel normalization in the study of natural speech," in *Linguistic Variation: Models and Methods*, edited by D. Sankoff (Academic, NY).

Jagharghi, A. J. (**1991**). "A comparative study of spectral peaks versus global spectral shape as invariant acoustic cues for vowels," PhD dissertation, Old Dominion University, August, 1990.

Markel, J. D. (**1972**). "The Sift algorithm for fundamental frequency estimation," IEEE Trans. Audio Electroacoust. **20**, 367–377.

McCandless, S. S. (**1974**). "An Algorithm for automatic formant extraction using linear prediction spectra," IEEE Trans. ASSP-**22**, 135–141.

Miller, J. D. (**1989**). "Auditory-perceptual representation of the vowel," J. Acoust. Soc. Am. **85**, 2114–2134.

Nearey, T. (**1978**). *Phonetic Feature Systems for Vowels* (Indiana University Linguistics Club, Bloomington, IN).

Nearey, T. M. (**1989**). "Static, dynamic, and relational properties in vowel perception," J. Acoust. Soc. Am. **85**, 2088–2113.

Neuburg, E. P. (**1980**). "Frequency-axis warping to improve automatic word recognition," ICASSP **80**, 573–575.

Neuburg, E. P. (**1988**). "Frequency warping by dynamic programming," ICASSP **88**, 573–575.

Nossair, Z. B. (**1989**). "Dynamic spectral shape features as acoustic correlates for stop consonants," PhD dissertation, Old Dominion University.

Oppenheim, A. V., and Johnson, D. H. (**1972**). "Discrete representation of signals," Proc. IEEE **60**(6), 681–691.

Paliwal, K. K., and Ainsworth, W. A. (**1985**). "Dynamic frequency warping for speaker adaption in automatic speech recognition," J. Phon. **13**, 123–134.

Syrdal, A. K. (**1985**). "Aspects of a model of the auditory representation of American English vowels," Speech Commun. **4**, 121–135.

Syrdal A. K., and Gopal, H. S. (**1986**). "A perceptual model of vowel recognition based on the auditory representation of American English vowels," J. Acoust. Soc. Am. **79**, 1086–1100.

Tanaka, K. (**1981**). "A parametric representation and clustering method for phoneme recognition—application to stops in a CV environment," IEEE Trans. Acoust., Speech, Sig. Proc. **29**, 1117–1127.

Wakita, H. (**1977**). "Normalization of vowels by vocal tract and its application to vowel identification," IEEE Trans ASSP **25**, 183–192.

Zahorian, S. A., and Gordy, P. E. (**1983**). "Finite impulse response (FIR) filters for speech analysis and synthesis," ICASSP **83**, 808–811.

Zahorian, S. A., and Jagharghi, A. J. (**1991**). "Minimum mean-square error transformations categorical data to target positions," IEEE Trans. ASSP. **39** (to be published).