# TRAINING MACHINE CLASSIFIERS TO MATCH THE PERFORMANCE OF HUMAN LISTENERS IN A NATURAL VOWEL CLASSIFICATION TASK

*Martin Hunke and Thomas Holton*

School of Engineering
San Francisco State University
San Francisco, CA 94132

## ABSTRACT

The purpose of this research is to determine how models of human auditory physiology can improve the performance of automatic speech recognition systems. In this study, a series of experiments was undertaken to discover how humans categorize and confuse vowels in natural speech. The recognition task comprised a large number of vowel nuclei isolated from naturally spoken sentences of a large number of talkers. Machine vowel classifiers were trained to match the results of these vowel categorization experiments using two input feature representations: a spectral–energy feature representation, and a representation derived from an auditory model. Classifiers trained to input representations derived from the auditory model match human performance and are more robust in the presence of noise and spectral filtering than classifiers trained to spectral–energy representations.

## 1. INTRODUCTION

Most current automatic speech recognition systems are based on a spectral–energy approach to feature extraction, such as computation of FFTs, LPC or cepstral coefficients. The performance of these systems is often severely degraded in noise and in conditions of spectral distortion ([1]). Signal processing approaches based on the human auditory system have been proposed to improve performance under these adverse conditions ([2]). In this work, we concentrate on the problem of vowel perception by human listeners and by machine classifiers and address two questions: 1) how do humans categorize and confuse vowels in natural speech? 2) can machine vowel classifiers be trained to match human performance on a natural vowel classification task?

## 2. METHODS

**Vowel database:** A corpus of 1845 steady–state vowel nuclei with durations ranging from 50—177 msecs was manually extracted from the TIMIT data-base. Nuclei occurred in CVC context in sentences spoken by 100 talkers (74 male and 26 female) in West Coast American English. The TIMIT–supplied phonetic labeling of sentences was used to establish the CVC context, but was otherwise not used in this study. Vowels were resampled at 8 kHz, normalized to a standard RMS amplitude and stored in disk files for use in the listening and machine classification experiments.

**Subjects:** Eight linguistically sophisticated listeners, 1 male and 7 female native speakers of West Coast English with ages ranging from 20 to 26, were recruited from undergraduate and graduate students in linguistics.

**Experimental protocol:** A computer–controlled experimental system presented listeners with sounds and recorded their responses. Sets of 100 vowel nuclei were selected by the computer at random from a universe of 1845 vowels, and each set was paired with an orthographic transcription of one of nine monothong vowel categories (/iy/, /ih/, /eh/, /ae/, /ah/, /aa/, /ao/, /uh/, and /uw/), displayed on the computer screen. Listeners rated each vowel's correspondence to the displayed vowel category on a scale of 0-3 by depressing appropriate keys on a keypad. Subsequent sets of 100 vowels was paired with a different displayed vowel category until listeners had judged all 1845 utterances in each of 9 vowel categories, producing a *raw response matrix (RRM)* of $1845 \times 9$ elements. Completing the entire experiment took listeners approximately 12 hours spread over many sessions, each of which was generally about one-half hour long.

In preliminary experiments, we determined that the most discriminating listeners assigned only 5.7% of vowel nuclei to the sum of categories 2 and 3 combined, whereas the least discriminating listeners assigned as much as 18.4% to these two categories. Because the analysis of data would have been confounded by averaging responses from individuals with markedly different overall standards of discrimination, we provided listeners with the guideline that only approximately 5% of utterances should be rated 2 or 3. Before proceeding with the main experimental task of 1845 vowels, listeners were screened with a shorter qualification task comprising only 100 vowels $\times$ 9 categories. This qualification task served to accustomize listeners to the experimental paradigm, and allowed us to assess their standard of discrimination and the consistency of their results. Three of the original 11 listeners recruited for these experiments were excluded from the main study based on their performance on the qualification task.

## 3. RESULTS

### 3.1. Evaluation of Listening Experiments

**Raw response matrices:** The percentage of ratings assigned by listeners to the four rating categories — 0, 1, 2 and 3 — varied

much more than the percentage assigned in the sum of categories. Accordingly, in our data analysis, we first coalesced the listener's responses into two binary categories: *n*, comprising ratings 0 and 1, and *y*, comprising ratings 2 and 3. The percentage of *y*–ratings from the 8 listeners ranged from 4.0% to 6.1%, and averaged 4.5%. Summarizing the data from these listening experiments, we find that:

*1) Individual listeners differ strikingly in their absolute classification of vowels.* Listeners classified few of the presented vowels (36%) as acceptable examples of *any* vowel category and classified only 30% unambiguously. There was also a substantial difference in categorical perception between listeners, even though these listeners all had similar standards of discrimination. Only 21% of vowels were classified by all listeners as belonging to no category at all, and only 2% of vowels were assigned to a single category on which all listeners agreed. Also, listeners differed significantly in the distribution of ratings they assigned to the 9 vowel categories; for example, the percentage of *y*–ratings assigned to /uw/ varied from 2% to 9%.

These data lead us to suggest that the natural listening task posed by these experiments, in which listeners are presented with a large number of vowel nuclei excised from natural running speech from a large number of speakers, is quantitatively different than the task posed by most vowel classification experiments, in which a small corpus of carefully produced vowels exemplars are drawn from clearly defined vowel categories are spoken in isolation or in highly constrained sentence contexts by a limited number of speakers ([4]). The fact that so few vowels in our experiment using natural speech can be classified unambiguously by themselves, buttresses the suggestion that the consonantal context in which these vowels appear in natural speech most likely plays a significant role in the natural speech task.

*2) Listeners make similar confusions in assigning vowels to vowel categories.* Although classification of vowels by individual listeners differ significantly, the confusions made between vowel classes are similar, and can be used to train and test machine vowel classifiers.

The confusions made by listeners in vowel classification are summarized by the covariance matrix. Each RRM, $M = (m_{ij})$, gives the rating of vowel $i$ for vowel class $j$, where $j$=0 (a rating of 'n') or $j$=1 (a rating of 'y'), The element $v_{ij}$ of the covariance matrix, $V = M^T M$, gives the number of vowels assigned to class $i$ that are also assigned to class $j$. Normalizing the covariance matrix by the percentages of vowels assigned to each of the vowel classes yields the *correlation matrix*, $CM = (c_{ij})$, where $c_{ij} = v_{ij}/v_{jj}$. To measure the difference between two CMs, we defined a measure, the *correlation distance*, $D_{\mathrm{CM}}$, which is the absolute difference between two CMs weighted by a factor related to the probability of vowel occurence:

$$D_{\mathrm{CM}}(R,S) = \sum_i \sum_j |r_{ij} - s_{ij}| \cdot p(j),$$

where $R$ and $S$ are the two matrices, and the weighting factor,

$$p(j) = \frac{v_{jj}}{\sum_k v_{kk}},$$

is computed from the covariance matrix corresponding to $R$. Because the weighting factor is computed solely from the data in $R$, this matrix is designated the *reference matrix*. The use of a reference matrix was motivated by the difficulty of comparing CMs under certain circumstances; specifically, RRMs generated by machine vowel classifiers, especially for vowels in noise and spectral distortion, often show degeneration, that is, they show few or no decisions in certain vowel classes, whereas humans show decisions in all vowel classes. The specification of a reference matrix and weighting factor prevents the correlation distance of degenerate CMs from being artifactually small.

CMs of all listeners were very similar and did not differ more from each other than CMs computed from the two halves of the RRM of a single listener. However, the CM of a hypothetical average listener, computed from the average of all RRMs, *did* differ from the CMs of individual listeners more than these CMs differed from each other. Accordingly, we chose the reference CM, $\overline{R}$, to be the average of the listeners' CMs (Table 1).

| | aa | iy | eh | ih | ae | ah | ao | uh | uw |
|---|---|---|---|---|---|---|---|---|---|
| aa | 100 | | | | | 7 | 67 | | |
| iy | | 100 | | | | | | | |
| eh | | | 100 | 2 | 6 | 3 | | 1 | |
| ih | | 1 | 2 | 100 | | | | 2 | 1 |
| ae | | | 10 | | 100 | 2 | | | |
| ah | 4 | | 2 | | 1 | 100 | 3 | 5 | |
| ao | 60 | | | | | 4 | 100 | | |
| uh | | | 1 | 1 | | 6 | | 100 | 7 |
| uw | | | | 1 | | | | 4 | 100 |

**Table 1:** $\overline{R}$, the average CM of all listeners, with numbers expressed as percentages.

## 3.2. Feature Representation

We compared the performance of multi-layer perceptron (MLP) vowel classifiers trained to two representations of speech: a conventional spectrographic representation and a representation based on a model of signal processing by the peripheral auditory system ([3]).

**Spectrogram:** Spectrograms were obtained by averaging the magnitude of the fast Fourier transform (FFT) of 256-point, Hamming-windowed frames of steady-state vowel nuclei (about 32 msecs), with subsequent frames overlapping by 128-points. 120 frequency points of this averaged FFT were retained, linearly spanning the frequency range 250 to 3974 Hz. Figure 1*A* shows the spectrogram of the vowel nucleus /eh/ in the absence of noise, in noise with S/N =+6 dB, and filtered by a 1-pole low-pass filter with a cut–off frequency of 250 Hz. Since the spectrogram is simply the frequency-domain, spectral–energy profile of sounds, this representation is sensitive to noise or spectral distortion of the input.

**Auditory model:** We have previously demonstrated an approach to the detection and categorization of important speech features in sonorant speech such as pitch pulses and formants based on a model of the peripheral auditory system ([3]). This approach is based on noting that the response of the auditory model to sonorant speech comprises two distinct epochs occurring in alternation, an *impulsive epoch*, in which auditory-nerve fibers throughout the cochlea respond at their characteristic frequency to a glottal pulse, and a

*synchronous epoch*, in which selected groups of fibers respond synchronously at the frequency of proximal formants. A formant is said to occur at locations along the cochlea and times at which an impulsive epoch is followed by a synchronous epoch. Figure 1*B* shows the time-averaged auditory–model representation to a steady-state vowel nuclei, /eh/. This representation, which comprises 120 channels spanning the frequency range from 250 to 3400 Hz, is quite sparse; in the absence of noise, there is substantial information only around frequencies that correspond to the formants. Because the auditory–model representation is based on a time-domain signal-processing strategy, it appears relatively insensitive to additive noise and spectral distortion of the input.

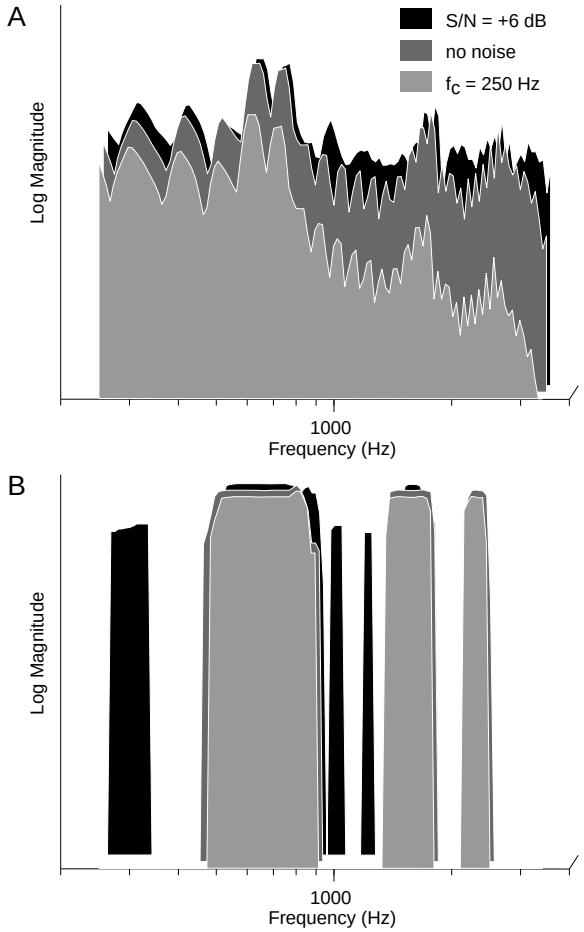


**Figure 1:** Magnitude of spectral representation (*A*) and auditory–model representation (*B*) to a vowel nuclei, /eh/, 70 ms in duration, in the absence of noise, in noise of S/N=+6 dB and filtering by a single pole infinite–impulse–response filter with cut–off frequency, $F_C$=250 Hz.

## 3.3. Machine Vowel Classification

**Architecture:** We employed MLP classifiers which had either the auditory–model or the spectral–feature representation of a vowel set as the input and generated an RRM as the output. The network architecture consisted of an input layer of 120 input units for the chosen feature representation, a hidden layer of 20 units, and an output layer of 9 units corresponding to the vowel classes. Larger hidden layers did not substantially increase network performance while fewer hidden units resulted in a performance deterioration.

**Training:** Networks were trained using the back–propagation method. Because the CM computed from the average of the RRMs of all listeners differed from the individual CMs, we trained networks to match individual listeners rather than to match a hypothetical average listener.

In order to train and test the classifiers, we split the complete set of 1845 vowels into three subsets, a training set (50%), a validation set (25%), and a test set (25%), each of which had the same balance of male and female speakers as the complete set (29% female and 71% male). RRMs from each of eight listeners were also split into corresponding "training", "validation" and "test" submatrices. During each training step, the RRM produced by a network was compared to the RRM derived from the "training" subset of a single human listener, and the network's weights were adjusted to reduce the Euclidean distance between the network's RRM and the listener's RRM. To enhance the generalization capabilities of the network, training of MLPs is commonly carried out until the Euclidean distance reaches a minimum on an independent validation set. However, because the CMs of different listeners appear more similar to each other than the RRMs, we trained the networks until $D_{\mathrm{CM}}$ between the CM of the network RRM and the CM of the reference matrix $\overline{R}$ reached a minimum on the validation set. For each listener, we trained 100 networks with different random initial weight settings and selected the 20 best networks, i.e. those with the lowest $D_{\mathrm{CM}}$. Computing a plurality of networks for each listener allowed us to eliminate networks that might have reached a local minimum of the Euclidean distance, and also allowed us to estimate the variance of network performance.

**Testing:** The performance of the classifiers was tested on vowels of the independent test set. We assessed the performance of networks primarily by using $D_{\mathrm{CM}}$ derived from comparing RRMs generated by eight network classifiers to $\overline{R}$. Specifically, we compared the average CM of the eight RRMs generated by the "best" MLPs with $\overline{R}$, then the average CM of the eight RRMs generated by the "second best" MLPs, and so on, to arrive at an *average correlation distance*, $\mu$ and an estimate of the variance, $\sigma$, which serve as indicators of network performance based on confusions between vowel classes. As an additional measure of classifier performance, we computed the average number of *y*–decisions that the networks and the corresponding humans had in common.

## 3.4. Network Performance

**Absence of noise:** In the absence of noise, the performance of spectral–energy and auditory–model representations was similar. Using the spectral–energy feature representation we obtained $\mu$=0.16 and $\sigma$=0.02 compared to $\mu$=0.12 and $\sigma$=0.02 for the auditory–model representation. Both mean values are close to the average correlation distance between any two human listeners using vowels of only the test set ($\mu$=0.13, $\sigma$=0.04), showing that the confusions of vowel classes by the networks in the absence of

noise resemble the confusions made by listeners in a natural vowel classification task. Furthermore, the number of identical *y*–ratings between each pair of network output and RRM of the corresponding listener were similar to the number of decisions any two listeners had in common.

**Noise:** Gaussian noise of variable amplitude was added to the vowel nuclei of the test list. Figure 2*A* shows $\mu$ and $\sigma$ as a function of the S/N ratio of the stimuli. To evaluate the increase of $\mu$, we used a multivariate pooled t–test at a level of significance of $\alpha$=1% and determined that for the spectral–energy feature representation, network CMs at S/N ratios of less than approximately +15 dB differed significantly from the CMs in the absence of noise, whereas for the auditory–model feature representation, a similar change of network CMs was found for S/N ratios of less than +6 dB. Still, at a S/N ratio of +9 dB in both representations, the average number of y–ratings identical to the human listeners had deteriorated to 59% (auditory model) and 69% (spectrum) of the corresponding values in the absence of noise. However, as a whole, these data suggest that the confusions made by the MLP vowel classifier networks employing auditory–model features are more invariant in the presence of noise than the confusions made by the networks using spectral features.

**Spectral distortion:** Vowels of the test set were processed with a one–pole infinite impulse response (IIR), discrete–time low–pass filter of variable cut–off frequency, $F_C$. Figure 2*B* shows $\mu$ and $\sigma$ as a function of $F_C$. For the spectral–energy feature representation, a filter with a cut–off frequency as high as 2000Hz causes a significant increase in $\mu$, while in case of auditory–model representation, $\mu$ remains relatively constant for values of $F_C$ as low as 125 Hz. Using a multivariate pooled t–test at a level of significance of $\alpha$=1%, we determined that for the spectral–energy feature representation, the network CMs produced at $F_C$=2000 Hz differed significantly from the CMs in the absence of filtering, whereas for the auditory–model feature representation, the CMs did not change significantly until at least $F_C$=125 Hz. For the spectral–energy feature representation, the average number of y–ratings the classifiers had in common with the human listeners at $F_C$=250 Hz dropped to half of the identical decisions found in the absence of filtering. In contrast, for the auditory–model representation, the number of identical decisions to the human listeners at this frequency did not change.

## 4. CONCLUSION

The major purpose of this study was to determine whether machine vowel classifiers can be designed which match human performance in listening experiments on vowels occurring in natural speech. We find that machine vowel classifiers can be successfully constructed using either the auditory–model or the spectral–energy feature representation as inputs. In the absence of noise or spectral distortion, the performance of classifiers trained to either of these representations does not differ substantially more from the human listeners than the human listeners differ from each other. However, classifiers employing the auditory–model representation as the input are more invariant in the presence of noise and spectral distortion than classifiers based on a spectral–energy feature representation.

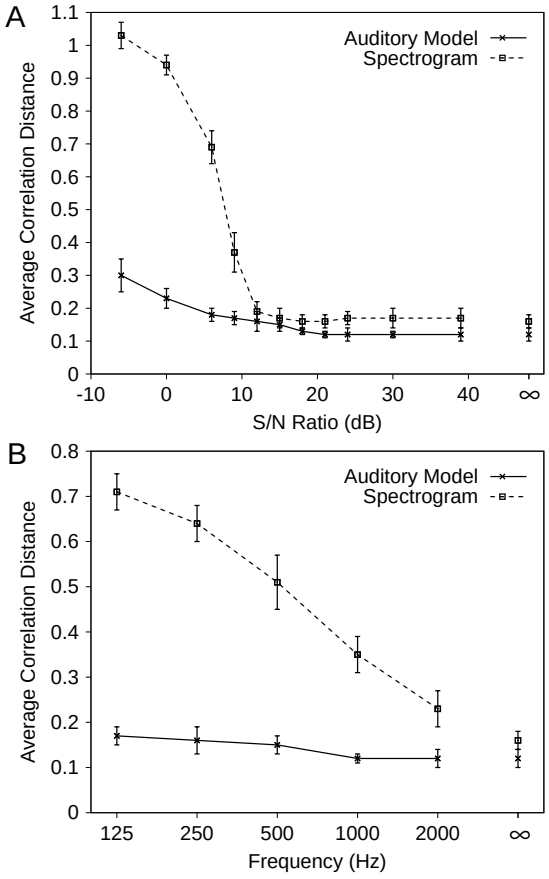


**Figure 2:** *A.* The average correlation distance, $\mu$, in additive Gaussian noise as a function of S/N ratio. *B.* The average correlation distance, $\mu$, in low–passed filtered speech, as a function of the filter's cut–off frequency, $F_C$.

## 5. REFERENCES


1. A. Acero and R. Stern. Environmental robustness in automatic speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 849–852, 1990.

2. O. Ghitza. Auditory nerve representation as a basis for speech processing. In S. Furui and M. Sondhi, editors, *Advances in Speech Signal Processing*, pages 453–485. Marcel Dekker, New York, 1992.

3. T. Holton and S.D. Love. Robust pitch and voicing detection using a model of auditory signal processing. *Proceedings of International Conference of Spoken Language Processing, Yokohama, Japan*, 1994.

4. W. Strange. Evolving theories of vowel perception. *J. Acoust. Soc. Am.*, 85:2081–2087, 1989.