Neural Network and Nearest Neighbor Comparison of Speaker Normalization Methods for Vowel Recognition

Gail A. Carpenter * and Krishna K. Govindarajan †
Center for Adaptive Systems and Department of Cognitive and Neural Systems
Boston University, 111 Cummington Street, Boston, MA, 02215-2411, USA

Abstract

Fuzzy ARTMAP and K-Nearest Neighbor (K-NN) categorizers were used to evaluate intrinsic and extrinsic speaker normalization methods by training and testing on disjoint sets of speakers of the Peterson-Barney database. Intrinsic methods included one nonscaled, four psychophysical scales (bark, bark with end-correction, mel, ERB), and three log scales, each tested on four different combinations of the frequencies F_0, F_1, F_2, F_3. Four extrinsic schemes were tested in conjunction with the intrinsic methods: centroid subtraction across all frequencies (CS), centroid subtraction for each frequency (CSi), linear scale (LS), and linear transformation (LT). Categorizers showed similar trends, with K-NN performing better but requiring more storage. The optimal intrinsic method was bark scale, or bark with end-correction, using differences between all frequencies (BDA). The order of performance for extrinsic methods was LT, CSi, LS, and CS, with ARTMAP performing best using BDA; and K-NN choosing psychophysical measures for all except CSi.

Speaker Normalization

Human listeners are able to identify as a single phoneme a wide variety of speech signals produced by different speakers in different contexts. For example, the vowel /æ/ is recognized despite the fact that the average F_1 formant frequency is 660 Hz for males and 1010 Hz for children [10]. *Speaker normalization* is a general term used to describe the process whereby a listener compensates for individual characteristics of a speech signal in order to extract invariant features needed to identify the sound.

This paper describes a procedure that can be used to make systematic comparisons of the many speaker normalization schemes that have been proposed in recent decades. To evaluate a given normalization method, the 1520 vowel token vectors of the Peterson and Barney (1952) database, each consisting of the fundamental frequency (F_0) and the first three formants (F_1, F_2, F_3), are preprocessed using that method. Normalized inputs from about 30% of the speakers (10 males, 9 females, and 5 children), corresponding to 480 vectors, are used to train three different classifiers, a neural network (fuzzy ARTMAP [3]) and two K-nearest neighbor systems[4]. The remaining test data set is then presented to each classifier, which tries to identify each as one of ten vowel sounds. The normalization scheme in question is evaluated in terms of the number of correct test set identifications made by each of the classifiers. Speaker independence is required since the test set inputs and the training set inputs are generated by disjoint sets of speakers. Comparative evaluations of 160 different normalization schemes were carried out using this method.

The two main classes of normalization methods are *intrinsic* and *extrinsic* [9]. Intrinsic normalization uses only the information present in each vowel token. Extrinsic normalization uses information from several vowel tokens of a given speaker. Intrinsic normalization methods include psychophysical measures, such as bark differences [11], logarithm measures [1, 7], and logarithms of formant ratios [7]. Extrinsic normalization methods include centroid subtraction across all frequencies (CS) [1, 9], centroid subtraction for each frequency (CSi) [1, 9], linear scale (LS) [6], and linear transformation (LT) [13].

We would like to thank R. Watrous for providing the Peterson-Barney database and for useful discussions on the linear transformation extrinsic method; and Ah-Hwee Tan for help with the fuzzy ARTMAP code.

*Supported in part by British Petroleum (BP-89-A-1204), DARPA (AFOSR-90-0083 and ONR-N00014-92-J-4015), the National Science Foundation (NSF- IRI-90-00530), and the Office of Naval Research (ONR-N00014-91-J-4100).

†Supported in part by the Air Force Office of Scientific Research (AFOSR-F49620-92-J-0225), DARPA (AFOSR-90-0083), and the National Science Foundation (NSF- IRI-90-00530).

Fuzzy ARTMAP and K-Nearest Neighbor Algorithms

Fuzzy ARTMAP [3] is a supervised neural network algorithm that learns to map (transformed) frequency vectors to vowel categories. ARTMAP clusters frequency vectors on-line in one module (ART_a) and vowel categories in a second module (ART_b). An intervening map field (F^{ab}) adaptively associates frequency categories to vowel categories. Performance was compared with that of K-nearest neighbor (K-NN) algorithms [4], using both city block (L_1) and Euclidean (L_2) metrics. The K-NN algorithm chooses a vowel category based on the K training points that lie nearest to a test point. Preliminary simulations on different normalization methods were used to choose parameters for the two different recognition methods. Fuzzy ARTMAP parameters for all the simulations were: $\bar{\rho}_a = 0.0$, $\alpha = 0.1$, and $\beta = 1.0$. For the K-NN systems, the number of neighbors (K) was fixed at 10 throughout.

Intrinsic Normalization Methods

For the intrinsic normalization schemes, eight normalization scales were compared: one nontransformed (N) scale; four psychophysical scales: bark scale (B) [15], bark scale with end-correction (Be) [12], mel scale (Mel) [5], and equivalent rectangular bandwidth scale (ERB) [8]; and three log measures: a semitone scale ($log_{1.06}$), natural log scale (log_e), and log base 10 scale (log_{10}).

The bark scale (B) transforms $F_0 \ldots F_3$ to $F'_0 \ldots F'_3$ according to the equation: $F'_i = 13.0 * \arctan(0.76 * F_i/1000) + 3.5 * \arctan(F_i/7500)^2$, where F_i is the i^{th} frequency, in Hz. Bark scale with end-correction (Be) adjusts the low frequencies before converting to them to bark scale: frequencies below 150 Hz are increased to 150 Hz; frequencies between 150 and 200 Hz are reduced to $0.8F_i + 30$; and frequencies between 200 and 250 Hz are increased to $1.2F_i - 50$. The mel scale (Mel) corresponds to the transformation: $F'_i = 2595 \log_{10}(1 + F_i/700)$. Finally, the equivalent rectangular bandwidth (ERB) scale is calculated by: $F'_i = 11.17 * \log_e((F_i + 312)/(F_i + 14675)) + 43$. The three logarithmic measures consist of the semitone scale, $F'_i = \log_{1.06}(F_i)$, the natural logarithm scale, $F'_i = \log_e(F_i)$, and the log base 10 scale, $F'_i = \log_{10}(F_i)$.

Each of the eight normalization scales was tested with four different combinations: only the first two formants $[F'_1, F'_2]$; the fundamental and all three formants $[F'_0, F'_1, F'_2, F'_3]$; the three differences $F'_1 - F'_0, F'_2 - F'_1, F'_3 - F'_2$ (Diff Subset); and all six difference combinations $F'_1 - F'_0, F'_2 - F'_0, F'_3 - F'_0, F'_2 - F'_1, F'_3 - F'_1, F'_3 - F'_2$ (Diff All). Syrdal and Gopal (1986) proposed the Diff Subset method with bark scale with end-correction. Nearey and colleagues [1, 9] and Miller and colleagues [7] proposed using log scaled frequency ratios, which correspond to the differences between the log coverted frequencies. Combining the 8 vowel space scales and the 4 frequency combinations, 32 intrinsic methods were tested.

Extrinsic Normalization Methods

For the extrinsic methods, adaptation to a speaker was superimposed on each of the 32 intrinsic normalization methods. Four types of extrinsic normalization were tested: centroid subtraction across frequencies (CS), centroid subtraction for each frequency (CSi), linear scale (LS), and linear transformation (LT). The CS method finds the mean frequency value ($\overline{F}$) across all transformed frequencies of all the vowels of a given speaker and subtracts this value from F'_i [1, 7, 9]. The CSi method extends the CS method by computing the centroid ($\overline{F_i}$) for each transformed frequency and subtracting this value from F'_i. The CLIH2 method [9], and CLIH3 method [1] are functionally equivalent to the CSi method in a log vowel space. The linear scale (LS) approach [6] finds the minimum and maximum frequency values for each F'_i across all vowels of a given speaker, then rescales each frequency to the range [0,999]. In the LT method [13], a linear transformation matrix $\mathcal{A}$ is obtained which transforms each speaker's frequencies into some prototypical frequency values. New frequencies are linear combinations of the original transformed frequencies: $F''_i = \sum_{k=0}^{3} \alpha_{ik} F'_k + \beta_i$. The matrix $\mathcal{A}$ is derived using the LMS algorithm [14] to minimize the mean squared error between a given speaker's fundamental and formant frequencies and the mean fundamental and formant frequencies across all speakers for each vowel. In all, 128 extrinsic normalization schemes were tested: 4 speaker adaptations x 4 frequency combinations x 8 scales.

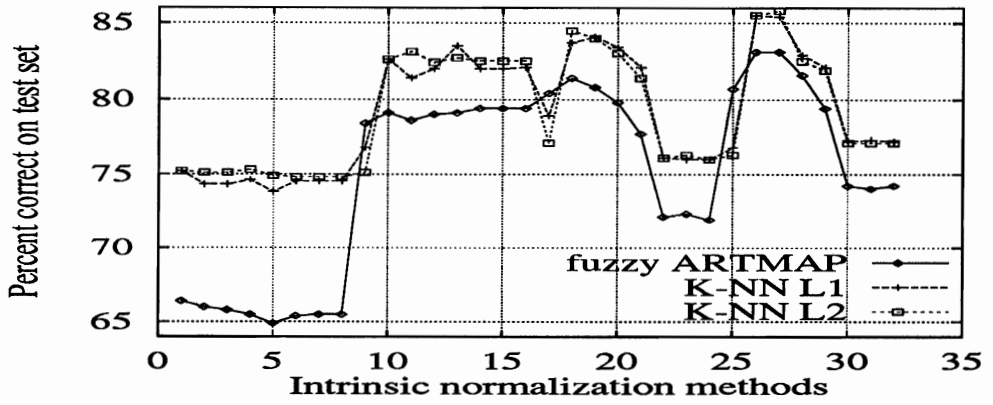

Figure 1: Test set performance of fuzzy ARTMAP and K-NN for intrinsic normalization methods # 1-32, which are identified in Table 1.

Comparative Evaluation of Normalization Methods

The three pattern recognition systems (fuzzy ARTMAP, L_1 K-NN, and L_2 K-NN) generally agreed on which normalization methods gave better predictive performance on test set data. K-NN tended to outperform fuzzy ARTMAP by a few percent (Figure 1). However, improved performance achieved by K-NN comes at a cost of storing all 480 training vectors. Fuzzy ARTMAP coded between 22 and 135 F_2^a nodes, providing a compression of 3.5 to 21.8 compared to the storage requirements of K-NN. Table 1 and Figure 1 show fuzzy ARTMAP and K-NN performance on the 32 intrinsic normalization methods. Similar analysis of the four extrinsic schemes has also been carried out [2].

Vowel Space	$[F_1, F_2]$			$[F_0, F_1, F_2, F_3]$			Diff Subset			Diff All		
	Fuzzy ARTMAP											
	Id	%	F_2^a	Id	%	F_2^a	Id	%	F_2^a	Id	%	F_2^a
N	1	66.4	123.1	9	78.4	63.5	17	80.4	55.8	25	80.7	57.5
B	2	66.0	123.7	10	79.1	61.6	18	81.4	56.3	26	**83.1**	43.9
Be	3	65.8	123.1	11	78.6	63.9	19	80.8	54.8	27	**83.1**	43.4
Mel	4	65.5	124.3	12	79.0	62.2	20	79.8	57.1	28	81.6	46.3
ERB	5	64.9	124.8	13	79.1	62.3	21	77.7	66.1	29	79.4	49.4
$log_{1.06}$	6	65.4	122.0	14	79.4	60.7	22	72.1	73.2	30	74.2	58.9
log_e	7	65.5	121.9	15	79.4	60.6	23	72.3	72.5	31	74.0	58.8
log_{10}	8	65.5	122.1	16	79.4	60.8	24	71.9	73.9	32	74.2	58.9
Vowel Space	K-NN											
	Id	L_1 %	L_2 %	Id	L_1 %	L_2 %	Id	L_1 %	L_2 %	Id	L_1 %	L_2 %
N	1	75.2	75.2	9	76.8	75.1	17	78.9	77.1	25	76.8	76.3
B	2	74.3	75.1	10	82.6	82.6	18	83.7	84.5	26	**85.5**	**85.5**
Be	3	74.3	75.1	11	81.4	83.1	19	84.1	84.0	27	**85.4**	**85.8**
Mel	4	74.6	75.3	12	82.0	82.4	20	83.4	83.0	28	82.9	82.5
ERB	5	73.8	74.9	13	83.5	82.7	21	82.1	81.4	29	82.1	81.9
$log_{1.06}$	6	74.5	74.8	14	82.0	82.5	22	76.1	76.1	30	77.2	77.1
log_e	7	74.5	74.8	15	82.0	82.5	23	76.0	76.3	31	77.3	77.1
log_{10}	8	74.5	74.8	16	82.1	82.5	24	76.0	76.0	32	77.2	77.1

Table 1: Fuzzy ARTMAP and K-NN test set performance with intrinsic normalization.

The psychophysical measures (B, Be, Mel, ERB) outperformed the log measures in most cases. For all the intrinsic and extrinsic methods, fuzzy ARTMAP performed best using bark, or bark with end correction, Diff All (Table 1). Although K-NN optimal performance varied more, these classi-

fiers also chose the psychophysical measures in all cases except for the extrinsic scheme CSi. For the intrinsic and LS extrinsic method, K-NN chose the bark Diff All method. For the CS extrinsic method, K-NN chose ERB $[F_0, F_1, F_2, F_3]$. For the LT method, L_1/L_2 K-NN performed best with Mel/ERB $[F_0, F_1, F_2, F_3]$. Finally, for the CSi method, K-NN chose the log scales $[F_0, F_1, F_2, F_3]$.

While the LT method has the best performance, it requires vowels that are labeled *a priori* to obtain the transformation matrix $\mathcal{A}$. Thus, for speaker-independent machine vowel recognition, LT requires the user to say an initial specified utterance containing the requisite vowels. The other three extrinsic methods do not require these vowels to be labeled. Thus, the second best method (CSi) may be the best candidate for prototype human and machine perpeption systems, since CSi does not require as much *a priori* knowledge as LT, its computational demands are less, and its performance is almost as good.

References

[1] Assmann, P. F., Nearey, T. M., and Hogan, J. T. (1982). "Vowel identification: Orthographic, perceptual and acoustic aspects," J. Acoust. Soc. Am. **71**, 975–989.

[2] Carpenter, G. A. and Govindarajan, K. K. (1993), "Speaker normalization methods for vowel recognition: Comparitive analysis using neural network and nearest neighbor classifiers," CAS/CNS Techical Report, Boston University, Boston, MA.

[3] Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. R., and Rosen, D. B. (1992). "Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps," IEEE Trans. on Neural Networks **3**, 698–713.

[4] Dasarathy, B. V. (1991). *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques* (IEEE Computer Society Press, Los Alamitos, CA).

[5] Fant, G. (1973). *Speech Sounds and Features* (MIT Press, Cambridge, MA).

[6] Gerstman, L. J. (1968). "Classification of self-normalized vowels," IEEE Trans. on Audio and Electroacoustics **AU-16**, 78–80.

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

[8] Moore, B. C. J. and Glasberg, B. R. (1983). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750–753.

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

[10] Peterson, G. and Barney, H. L. (1952). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

[11] 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.

[12] Traunmüller, H. (1981). "Perceptual dimension of openness in vowels," J. Acoust. Soc. Am. **69**, 1465–1475.

[13] Watrous, R. L. (1993). "Speaker normalization and adaptation using second-order connectionist networks," IEEE Trans. on Neural Networks, **4**, pp. 21–30.

[14] Widrow, B. and Stearns, S. D. (1985). *Adaptive Signal Processing* (Prentice Hall, Englewood Cliffs, NJ).

[15] Zwicker, E. and Terhardt, E. (1980). "Analytical expression for critical-band rate and critical bandwidth as a function of frequency," J. Acoust. Soc. Am. **68**, 1523–1525.