VOICE IDENTIFICATION USING NEAREST-NEIGHBOR DISTANCE MEASURE

A.L. Higgins, L.G. Bahler, and J.E. Porter

ITT Aerospace/Communications Division, San Diego, CA 92131 USA

1. ABSTRACT

An algorithm is described for attributing a sample of unconstrained speech to one of several known speakers. The algorithm is based on measurement of the similarity of distributions of features extracted from reference speech samples and from the sample to be attributed. The measure of feature distribution similarity employed is not based on any assumed form of the distributions involved. The theoretical basis of the algorithm is examined, and a plausible connection is shown to the divergence statistic of Kullback[Fuku72]. Experimental results are presented for the King telephone database and the Switchboard database.

2. INTRODUCTION

Speaker identification is generally accomplished by comparing an unattributed input utterance with reference utterances from each of several known speakers and attributing the input to the known speaker whose utterances are determined to be most similar to the input. For unconstrained speech utterances, a comparison method is required that is sensitive to voice differences but relatively insensitive to the phonetic content of the utterances. The method used, which we refer to as the nearest neighbor (NN) method, is defined in Section 3. The NN method is so termed because it is based on the measured distances from each frame of an utterance to the nearest other frame of the same utterance and to the nearest frame of each other utterance being compared.

The NN method was developed experimentally and found to give excellent performance. The explanation provided in Section 4 attempts to rationalize the algorithm in terms of probability density function (PDF) estimation and statistical pattern recognition. Algorithms based on PDF estimation have been shown to be superior to most "minimum distance" algorithms, which do not explicitly entail such estimation[Schw82]. Proceeding from a conjecture regarding the local relationship between probability density and nearest-neighbor distances, the NN algorithm is shown to measure global differences between the speakers' underlying feature distributions.

The complete speaker identification algorithm using the NN distance measure is described in Section 5. Pre-processing operations, designed to minimize sensitivity to noise and possible differences in channel frequency response, are an important part of the algorithm. Pre-processing includes an effective pruning mechanism to reduce memory and computation requirements, enabling practical applications of the algorithm to be realized using currently available processors. Performance of the algorithm is reported in Section 6, and conclusions are summarized in Section 7.

3. NEAREST NEIGHBOR DISTANCE MEASURE

We define the nearest neighbor distance, $d(U,R)$, between unknown utterance U and reference utterance R as follows:

$$\begin{aligned} d(U,R) \equiv\ & \frac{1}{|U|} \sum_{u_i \epsilon U} \min_{r_j \epsilon R} |u_i - r_j|^2 \\ & + \frac{1}{|R|} \sum_{r_j \epsilon R} \min_{u_i \epsilon U} |u_i - r_j|^2 \\ & - \frac{1}{|U|} \sum_{u_i \epsilon U} \min_{\substack{u_j \epsilon U \\ j \neq i}} |u_i - u_j|^2 \\ & - \frac{1}{|R|} \sum_{r_i \epsilon R} \min_{\substack{r_j \epsilon R \\ j \neq i}} |r_i - r_j|^2 \\ \equiv\ & e(U,R) + e(R,U) - e(U,U) - e(R,R). \end{aligned} \tag{1}$$

where $\{u_i\}$ and $\{r_i\}$ are feature vectors or frames belonging to U and R, and $|U|$ and $|R|$ are the number of frames in U and R, respectively. The basis of this measure and its interpretation in terms of classical pattern recognition is discussed in the next section.

4. MATHEMATICAL BASIS

In the discussion below we assume $R = \{r_i\}$ and $U = \{u_i\}$ are independent random samples from underlying distributions with continuous densities p_U and p_R, respectively. These distributions are ordinary probabilities on a feature (or parameter) space, **F**, which can be thought of as a Euclidean vector space of several (fourteen in this case) dimensions.

4.1. The Affine Connection

Use of the squared Euclidean metric for nearest neighbor classification procedures are rationalized in pattern recognition textbooks by the relationship between the Parzen estimate of local probability density and nearest neighbor distance[Duda73, Fuku72]. It is usually shown in such sources that nearest neighbor classification is equivalent to estimating the local density of each pattern class by a Parzen estimate and assigning the unknown to the class

0-7803-0946-4/93 $3.00 © 1993 IEEE

with the largest local density, as is required for a Bayesian decision (in the case of equal priors and equal costs). Generalizing to the case where many observations (assumed independent) are to be combined to reach a decision, the appropriate paradigm would be to estimate the log-likelihood ratio using Parzen density estimates and accumulate those estimates. Again following the textbooks, the Parzen estimate of the local density of a reference class at the observation point of the unknown is

$$p_R \approx \frac{1}{V_n(d_{NN})},$$

where $V_n(r)$ is the volume of a sphere of radius r in the n-dimensional feature space, and d_{NN} is the distance from the observation point to its nearest neighbor in the reference class R. Since $V_n(r)$ is proportional to r^n, to accumulate log likelihood ratios we should accumulate values proportional to

$$-\log(p_R) \approx n \log(d_{NN}).$$

Empirically we have found that accumulating the squares, rather than the logarithms, of nearest neighbor distances gives much better performance. This discrepancy was at first profoundly puzzling, since it is clearly best to accumulate a quantity which approximates the logarithm of the likelihood ratio, as the logic of Bayesean decision making incontrovertibly shows. The theory of Parzen estimators strongly indicates that log densities are much more closely related to the logarithm of nearest neighbor distances than to their square.

One possible explanation of this basic discrepancy between theory and experience is that a "tyranny of dimension" invalidates the application of simple Parzen-like arguments to the speaker recognition setting. Parzen estimation assumes that the density at the test point and at its nearest neighbor in the sample are about the same. In a feature space of five or more dimensions immense sample size - far beyond practical speech or speaker recognition limitations - is necessary for this assumption to be viable, even when the test point distribution is the same as that of the density being estimated.

There is a further possible reason for the discrepancy. One may hypothesize a circumstance under which the local log probability density is much better approximated by an affine function of the squared nearest neighbor distance than by its logarithm, and that hypothetical circumstance may quite accurately reflect important aspects of speech feature distributions and hence the speaker recognition problem. The hypothetical setting assumes a smooth, multi-dimensional distribution which has roughly Gaussian characteristics at large distances from a central accumulation. That is, the log density is assumed to fall off roughly as the square of distance from the distribution center. Consider what happens when one attempts to estimate the local density using nearest neighbor distance from a test point to a limited sample from this distribution. When the log density falls with the square of the distance from the center of the distribution, the samples will be concentrated near the distribution center. This can be appreciated by realizing that the distribution of squared distance of a random sample from the center will be like a χ^2 variate, with number of degrees of freedom equal to the dimensionality of the feature space. The χ^2 distribution falls off very rapidly on the right in five or more dimensions, so even enormous samples are confined within a limited distance of the distribution center. When the test point is at a large distance from the distribution center, the nearest neighbor will tend to be much nearer the distribution center, so the square of the nearest neighbor distance will tend to be roughly the squared distance from the test point to the distribution center. Test points at arbitrarily large distances from the distribution will thus have nearest neighbor squared distances which grow like the square of the distance of the test point from the distribution center. The log density is approximately an affine function of that squared distance, by virtue of the quasi-Gaussian nature of the hypothetical distribution, so the square is the appropriate function of that distance to use, especially for test points at great distance from the distribution center.

An argument can also be made that an affine function is better than the logarithm for test points near the center of the hypothetical distribution. The assumed density has an upper bound, by virtue of its quasi-Gaussian nature. The negative logarithm of the density then has a finite lower bound. However, nothing precludes arbitrarily small nearest neighbor distances from occurring, particularly when the test point is in the vicinity of the distribution mode. The logarithm of nearest neighbor distance therefore has no lower bound, unlike the negative logarithm of the density. Since small nearest neighbor distances may occur for any test point, it is more appropriate that the log density approximating function should have a finite limit as nearest neighbor distance approaches zero; an affine function of the squared distance has that property and the logarithm of that distance does not.

These reflections were suggested and verified by Monte Carlo simulation experiments using random samples of a wide range of sizes from a variety of different densities and feature space dimensionality. The simulations showed that the conditional expectation of the logarithm of local density, given the nearest neighbor distance, is approximately an affine function of that distance squared, and that this is true for a wide variety of different distribution shapes. We conjecture that this is a property of the distribution of the speech features used in our algorithm: *i.e.*,

$$-\ln(p_R(u \mid d_{NN})) \approx \alpha + \beta d_{NN}^2, \qquad (2)$$

for some constants α and β. Equivalently,

$$d_{NN}^2 \approx -\frac{1}{\beta}\left[\alpha + \ln(p_R(u \mid d_{NN}))\right] \qquad (3)$$

Testing this conjecture appears very difficult, as any conventional test would require an impractically large speech sample, in view of the high dimensionality (14) of the feature space.

4.2. Implications of The Affine Connection

Consider the first term, $e(U,R)$, in Equation 1. It is the average squared distance over all frames u_i in U to their nearest frames in R. Applying the affine connection as expressed in Equation 3,

$$e(U,R) \approx \frac{1}{|U|} \sum_{u_i \varepsilon U} \left\{ -\frac{1}{\beta_R} \left[\alpha_R + ln(p_R(u_i)) \right] \right\}$$

Here α and β have been written with subscripts as a reminder that they depend on the density, p_R, and the "size" of the sample, $|R|$. The approximation is unbiased if the sampling of p_R to make R and p_U to make U are independent events. The expected value of $e(U,R)$ with respect to these two random samplings is then just the expectation of the averaged quantity with respect to the test-point distribution, p_U. That is,

$$\begin{aligned} E[e(U,R)] &= E\left[-\frac{1}{\beta_R}[\alpha_R + \ln p_R]\right] \\ &= -\frac{1}{\beta_R}\int_{\mathbf{F}}\left[\alpha_R + \ln p_R\right]p_U df \\ &= \frac{1}{\beta_R}\int_{\mathbf{F}} p_U(-ln\ p_R)df - \frac{\alpha_R}{\beta_R} \end{aligned}$$

Analogous relations apply to the other terms in Equation 1. The expected value of the nearest neighbor distance is then

$$\begin{aligned} E[d(U,R)] &= \frac{1}{\beta_R}\int_{\mathbf{F}} p_U(-ln\ p_R)df \\ &+ \frac{1}{\beta_U}\int_{\mathbf{F}} p_R(-ln\ p_U)df \\ &- \frac{1}{\beta_R}\int_{\mathbf{F}} p_R(-ln\ p_R)df \\ &- \frac{1}{\beta_U}\int_{\mathbf{F}} p_U(-ln\ p_U)df \end{aligned} \tag{4}$$

The divergence, a classical measure of dissimilarity between probability distributions, when applied to p_U and p_R is defined as:

$$\mathbf{D}(p_R,p_U) \equiv \int_{\mathbf{F}}(p_R - p_U)[\ln(p_R) - \ln(p_U)]df$$

Writing out the integrand and integrating terms separately,

$$\begin{aligned} \mathbf{D}(p_R,p_U) &= \int_{\mathbf{F}} p_U(-ln\ p_R)df \\ &+ \int_{\mathbf{F}} p_R(-ln\ p_U)df \\ &- \int_{\mathbf{F}} p_R(-ln\ p_R)df \\ &- \int_{\mathbf{F}} p_U(-ln\ p_U)df \end{aligned} \tag{5}$$

We refer to the first two terms as *cross-entropies* and the second two terms as *self-entropies*. Note that the expected value of each term in Equation 1 is proportional to one of the entropy terms in Equation 5. Thus, the conjectured affine connection leads to a close relationship between the divergence and the expected value of the nearest neighbor distance measure, as shown by the similarity between equations 4 and 5. If $\beta_R = \beta_U = \beta$, then

$$\mathbf{D}(p_R,p_U) = \beta E[d(U,R)]$$

The divergence has many desirable properties as a measure of dissimilarity of distributions and it is at least plausible that the excellent performance of the NN algorithm derives from this (conjectured) close relationship to the divergence measure.

5. VOICE IDENTIFICATION ALGORITHM

The input utterance and the reference utterances are preprocessed identically in the manner described below. The pre-processed input utterance is then compared with each pre-processed reference utterance using Equation 1. The speaker of the reference message with the lowest nearest-neighbor score is identified. Pre-processing consists of: performing a filterbank analysis on the input sampled waveform; pruning low-amplitude frames; blind deconvolution; differencing with respect to frequency; and pruning redundant frames.

A 14-channel FIR filterbank computes spectra in Mel-spaced frequency bands at a rate of 50 frames per second. The spectra are compressed using the fourth-root function, and l_2 normalized to be insensitive to gain[Olan83].

Low amplitude frames are eliminated. The amplitude pruning threshold is established from a histogram of amplitude computed over all frames in the utterance. The threshold is set equal to the 10-*th* percentile of the histogram, plus 6 dB.

Channel effects are minimized by deconvolving the spectral data so that the mean value of each filterbank channel (the fourth root of power) over the selected frames of each utterance is equal to a constant. The l_2 normalization is re-applied to each frame after deconvolution.

Within each frame, differences are computed between adjacent frequency channels:

$$s'_i = \begin{cases} s_0 - s_{13} & i=0 \\ s_i - s_{i-1} & 0<i<14 \end{cases}$$

where s_i is the normalized fourth root of power in the i-*th* filterbank channel, after deconvolution.

Finally, redundant frames are discarded by sequentially testing each frame to determine whether it is within a pre-defined distance threshold of any frame already kept. Typically the distance threshold is set equal to the average value of cross entropy between utterances of the same speaker. This step provides a large reduction in the required computation and memory with minimal effect on performance.

6. PERFORMANCE

6.1. Databases

The algorithm was tested on the King telephone database and on the Switchboard database. The King database contains ten sessions from each of 51 male speakers. Each session is about 30-40 seconds in duration, and contains speech from a single speaker that is unconstrained except for the general topic. A change of recording equipment was made during collection of the database, introducing a channel difference between the first and second half (roughly between sessions 5 and 6). To observe the effect of this channel difference, tests were performed using training and test data from the same half and from different halves.

The Switchboard database was described by Godfrey, et. al[Godf92]. Because of the large size of the database, only a portion of it was used in the testing. 24 speakers (12 male and 12 female) were selected. Mark files were used to restrict recognition to the time intervals when one designated person (and not that person's conversation partner) was talking.

6.2. Test Description

Closed-set identification tests were performed, wherein the "unknown" test speaker was always one of the known reference speakers. Performance was measured by percentage correct recognition and by the average rank of the correct speaker among all reference speakers. (1.0 corresponding to no errors.) In all cases three sessions of each speaker were used for training. The result labeled "King (same channel)" in the table below is the average obtained from training on sessions 1, 2, and 3 and testing on sessions 4 and 5, and training on sessions 6, 7, and 8 and testing on sessions 9 and 10. The result labeled "King (cross channel)" is the average obtained from training on sessions 1, 2, and 3 and testing on sessions 9 and 10, and training on sessions 6, 7, and 8 and testing on sessions 4 and 5.

6.3. Results

Database	Percent Correct	Average Rank
King (same channel)	79.9	2.06
King (cross channel)	68.1	2.62
Switchboard	95.9	1.07

7. SUMMARY

The NN distance measure was defined, together with other elements of a text-independent speaker identification algorithm. Performance of the algorithm is better than that reported for algorithms based on Gaussian modeling and robust discrimination[Gish90]. An algorithm similar to the NN algorithm is the vector quantization (VQ) algorithm of Rosenburg and Soong[Soon85]. Results reported elsewhere in this proceedings[Kao93] show that the NN and VQ algorithms in fact give comparable performance on the King database. An argument was presented that the NN distance measure is related to the divergence, based on a conjectured relationship (the "affine connection") between speech feature probability density and nearest neighbor distances. The conjectured relationship is difficult to verify directly, as the required speech sample size is impractically large.

References

Duda73.
R. Duda and P. Hart, *Pattern Classification and Scene Analysis,* Wiley and Sons, Inc., 1973.

Fuku72.
Keinosuke Fukunaga, *Introduction to Statistical Pattern recognition,* Academic Press, New York, New York, 1972.

Gish90.
Herbert Gish, "Robust discrimination in automatic speaker identification," *Proc. Internatl. Conf. Acoust., Speech, and Sig. Proc.*, pp. 289-292, 1990.

Godf92.
Godfrey, J., Holliman, E., and McDaniel, J., "Switchboard: Telephone Speech Corpus for Research and Development," *ICASSP-92*, vol. 1, pp. 517-520, San Fransisco, March 1992.

Kao93.
Y.H. Kao, P.K. Rajasekaran, and J.S. Baras, "Robust Free-Text Speaker Identification Over Long Distance Telephone Channels," *ICASSP-93*, Minneapolis, April 1993.

Olan83.
C. Olano, "An investigation of spectral match statistics using a phonetically marked data base," *IEEE Internat. Conf. Record on Acoust., Speech and Signal Process.*, no. ICASSP-83, 1983.

Schw82.
R. Schwartz, S. Roucos, and M. Berouti, "The application of probability density estimation to text-independent speaker identification," *Proc. IEEE Internat. Conf. Acoust., Speech and Signal Process.*, pp. 1649-1652, 1982.

Soon85.
F. Soong, A. Rosenberg, L. Rabiner, and B. Juang, "A Vector Quantization Approach to Speaker Recognition," *Proc. ICASSP 85*, vol. 1, pp. 387-390, Tampa, FL, 1985.