A Text Independent Speaker Recognition System Using a Novel Parametric Neural Network

Paul Gomez, Ph.D.

Quantum Nanoelectronics, Inc., R&D Department, 12851 Luray Rd, SW Ranches, FL 33330, USA, Phone: 1-954-536-7326, Email: pgomez@qnanotronics.com

Abstract

This paper presents a new Speaker Recognition Technique aimed at high identification accuracy and low impostor acceptance. This method is based on a modified neural network, which is an extended and improved version of a Self-Organizing Map in multiple dimensions. The goal of this methodology is to achieve high accuracy identification and impostor rejection. The proposed method, Multiple Parametric Self-Organizing Maps (M-PSOM) is a classification and verification technique. This novel method was successfully implemented and tested using the CSLU Speaker Recognition Corpora of the Oregon School of Engineering with excellent results. This method builds a unique parametric neural network for each speaker as opposed to a single neural network for the whole system as it has been done in the past. With this technology a parametric neural network is a unique representation of a speaker's acoustic signature.

Keywords*: Speaker Recognition, SOM, PSOM, Forensic Science, Neural Networks, Speech Processing*

1. Introduction

Accuracy is very important in all applications but particularly in forensic science, which is the motivation of this research. Audio recordings are commonly presented in Courts of Law as evidence and without an adequate automated way to verify identity there will always be room for doubt. Although a machine will never determine with 100% accuracy the identity of a person based on his or her voice the same can be said of human beings. Nevertheless, a machine will yield a result that is an adequate point of reference that is not subjective or biased in any way.

Speaker Recognition is the process of automatically recognizing a person who is speaking on the basis of individual parameters included in his/her voice. This technology allows information systems to use the user's voice to verify identity and control access to services such as banking by telephone and many other applications. In forensic science this technology could be used to determine if a recorded voice belongs to a particular subject.

Most of the technologies used for Speaker Recognition have been borrowed from Speech Recognition, since this is a mature discipline and many of the concepts and methods can be readily applied. However, in Speech Recognition the objective is to recognize the words being spoken regardless of the speaker, whereas in Speaker Recognition the goal is to recognize the speaker regardless of the words being uttered.

A Speaker Recognition System has the following two main modules: 1) Feature Extraction and 2) Classification and Verification. For feature extraction the most commonly used techniques are MEL-Frequency Cepstrum Coefficients (MFCC) and Linear Predictive Coding

(LPC) [1-4], Wavelets [5-11] and other specialized techniques [12-20]. For classification and verification, technologies such as Vector Quantization (VQ) [21-22], Hidden Markov Models (HMM & GMM) [11, 23-27] and Neural Networks have been used [8, 15, 28-29].

2. The Multiple Parametric Self-Organizing Maps (M-PSOM)

The SOM neural network was invented by Kohonen [30-31] to classify simple patterns, such as letters fonts consisting of arrays of pixels, where a single input vector *x* represents a single pattern. In this way, several representations (fonts) of the same letter are presented to the neural network during several epochs until eventually the SOM allocates a unique cluster for every single class (a letter, in this example).

For Speaker Recognition, the problem is somewhat different and more complex because the speech signal recorded from a speaker utterances, contains multiple acoustic vectors that represent a spoken sentence and therefore, the simplified single-vector input cannot be used as a model of the subject's voice and consequently a single SOM cluster cannot be designated to represent a speaker.

Specifically in the experiments performed in this research, a single utterance of 5 digits may contain between 200 and 500 acoustic vectors and a training session consisting of multiple sentences, contains around 2000 acoustic vectors per speaker. An acoustic vector is a set of numbers that represent the power of the speech signal in different frequency bands during a short time frame in the order of 20 to 30 ms. As it can be seen, representing a speaker is not just a matter of clustering a few input vectors that define the same pattern, but rather, the problem is to create an acoustic space of *M* clusters for a single speaker such that the SOM in this way arranged, will uniquely identify a single speaker. The SOM trained in this manner will become the speaker's *acoustic signature*.

Another drawback of the basic SOM is that it does not contain any additional information that would allow us to reject an impostor. In the example presented before, if a letter that is not part of the training set (an impostor) is presented to the SOM, the result will be the letter with the most similar shape which corresponds to the cluster that is closest in distance to the sample pattern. Thus, an erroneous character is identified. The basic SOM model was designed for identification by means of minimum distance, but not for verification.

3. The Parametric SOM (PSOM) Model

Based on our previous discussion, it is necessary to add information to the basic SOM to be able to discriminate a pattern, once it has been associated with a cluster. This additional information consists of parameters that represent the size and density of the cluster. The following figure shows a 2-dimension cluster with its main parameters (cluster *j*, dimensions *i* and *k*).

The parameters added to each cluster are:

- μ, the average distance from all vectors clustered together to the centroid of the cluster
- σ, the standard deviation of the distances from all vectors to the centroid.
- ρ, the radius or size of the cluster. It is the maximum distance found from all vectors to the centroid of the cluster.
- d, the density, represents the number of vectors associated with a cluster

- β, the extended radius of acceptance. It's greater than ρ by 30% and it was found experimentally.

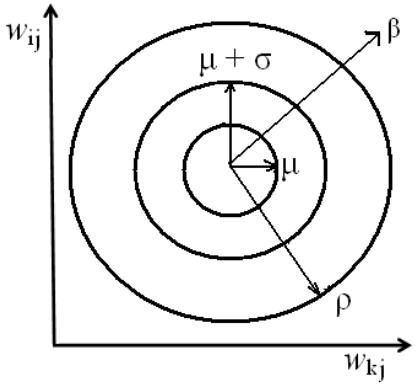

Figure 1 SOM Cluster Parameters

To understand how these parameters will improve the accuracy, let's refer to Figure 2, showing a Parametric SOM that has 3 clusters.

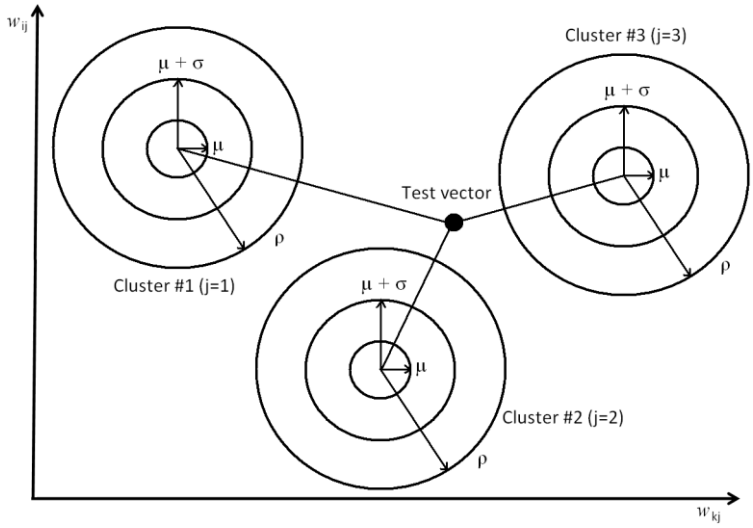

Figure 2. Example of a Parametric SOM

In this example, it can be seen that the nearest cluster to the test vector is cluster number 2. However, if the rule now is that the distance between a test vector and its nearest cluster has to be smaller than the radius of the cluster, this test vector will be rejected. With this simple scheme, several validity rules can be derived, depending on the application. One rule could be to reject vectors whose distances fall beyond μ+σ. It was found experimentally that approximately 30% of a speaker's test acoustic vectors fall in this range. This new threshold for vector acceptance is called β. Therefore, the most flexible rule was used, that is, to reject vectors whose distance is greater than ρ. It was also found experimentally, that when comparing a valid speaker test speech against its trained PSOM, up to 7.5% of the acoustic vectors fall beyond the boundaries of any cluster. This result is used in the algorithm for identification and verification.

However, this rule alone is not sufficient to verify a speaker. Since the SOM algorithm performs its classification by measuring the Euclidian distance from test vectors to all its clusters centroids, it is also necessary to find the global distance from all acoustic vectors to all clusters. The exact algorithm will be presented in the next section. First, let's introduce the new parametric SOM model, shown in Figure 3.

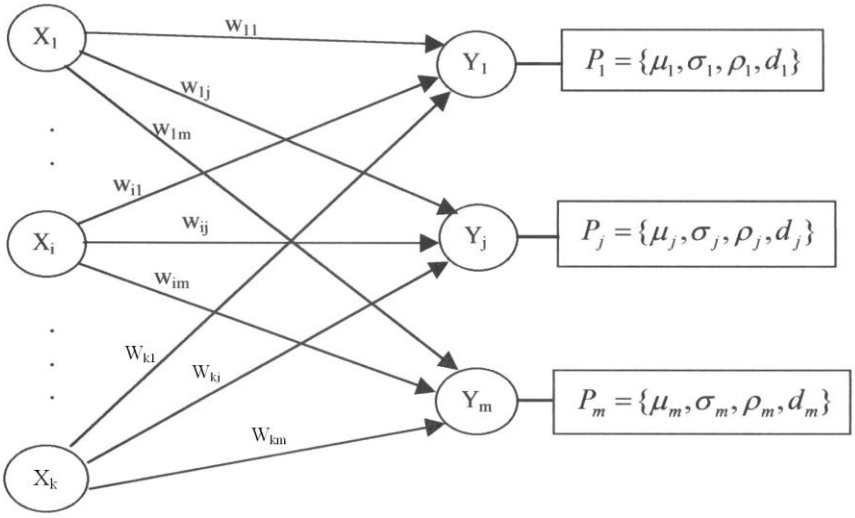

Figure 3. Parametric SOM Architecture

The parameters of the PSOM can be stored in a matrix ***P*** of *M* rows (one per cluster) and 4 columns, one for each parameter. Not all parameters are required for all applications. Particularly, the standard deviation, σ_i, requires an additional cycle of computations since it depends on the mean, μ_i. If the standard deviation is not required, it should be left out of the model to improve performance.

4. The MPSOM Training Algorithm

The algorithm is tailored to Speaker Recognition and in this system a single PSOM represents a single Speaker, that is, a PSOM is the user's unique *acoustic signature*. Thus, for multiple Speakers we will use Multiple PSOM (MPSOM). The input to the algorithm is the set of all Acoustic Vectors (***AV***) obtained after feature extraction performed with a MFCC processor. The following are the steps of the training algorithm.

Step 0. Initialization of weights:

- All the acoustic vectors (***AV***) are collected first, *N* vectors of *K* bands each.
- Weight matrix **W** with w_{ij}, i=1, .. K, j=1, .. M, is initialized with the histogram frequency distribution of each band *K* of ***AV***.
- Values of each band *K* are statistically distributed into *M* bins and stored into the vector ***h***.
- Each row of **W** is initialized with the corresponding vector ***h***.

Step 1. Set topological neighborhood parameters (linear, initial radius = 2).

- Set learning rate parameters ($\alpha = 0.5$, decreasing linearly to 0.1).
- Initialize the parameters matrix $\boldsymbol{P}(M,4)$ with zeroes.
- Initialize the matrix of final distances $\boldsymbol{FD}(M,N)$ with zeroes.

Step 2. Perform steps 3-9 E times, where E is the maximum number of epochs ($E = 10$ was found to be an adequate value)

Step 3. For each input vector **x**, do steps 4-7.

Step 4. For each cluster j, we compute:

$$D(j) = \sum_i (w_{ij} - x_i)^2$$

Step 5. Find index j^* such that $D(j^*)$ is a minimum and calculate its corresponding distance Z.

Step 6. For all units j within a specified neighborhood of j^*, and for all i, update their weights:

$$w_{ij}(new) = w_{ij}(old) + \alpha[x_i - w_{ij}(old)]$$

Step 7. On the last cycle (i.e., E = maximum) do:
Increment cluster density and store it:

$d_j = d_j + 1$
$\boldsymbol{P}(j^*, 4) = d_j$
Store distance:
$\boldsymbol{FD}(j^*, d_j) = Z$

Step 8. Update learning rate (α).

Step 9. Initial radius is 2. After half of the epochs are processed, the radius is reduced to 1, and 0 when ¾ of the epochs are reached.

Step10. Compute statistical parameters:
For cluster j=1, .. M do:
If density $d_j > 0$ then
Let s = {a set of distances in $\boldsymbol{FD}$ from 1 to d_j}

$$\mu_j = \text{average}(s) = \frac{1}{d_j}\sum_{k=1}^{d_j} s_k$$

$$\sigma_j = \sqrt{\frac{1}{d_j}\sum_{k=1}^{d_j}(s_k - \mu_k)^2}$$

$\rho_j = \max(s)$
$\boldsymbol{P}(j,1) = \mu_j$
$\boldsymbol{P}(j,2) = \sigma_j$
$\boldsymbol{P}(j,3) = \rho_j$
end if
end for

The initialization of the weights in step 0 is done with the known statistics of the acoustic vectors of the speaker to be clustered, instead of using random numbers, which is the traditional method. The training of the MPSOM was found to be faster with this initialization technique, requiring only ten epochs (E = 10) for training, while with random numbers takes up to 10 times longer.

5. The MPSOM Architecture

In this architecture, the recorded voice of each Speaker in digital format is segmented in time frames corresponding to 30 ms of the speech signal. Each frame is processed by a MFCC Feature Extraction Processor to obtain an acoustic vector for each frame. Thus, every speaker has a set of training vectors (acoustic vectors), ***AV*** that is used to train its PSOM model. Each acoustic vector has 16 values corresponding to the power in 16 MFCC frequency bands as shown schematically in Figure 4. After training, the system will have Multiple PSOM models, one for each trained speaker. Figure 5 shows the new proposed architecture for U Speakers. As more users are added to the system, the number of PSOMs will increase but no retraining of the previous PSOMS is necessary which results in a great advantage of this method compared to other solutions.

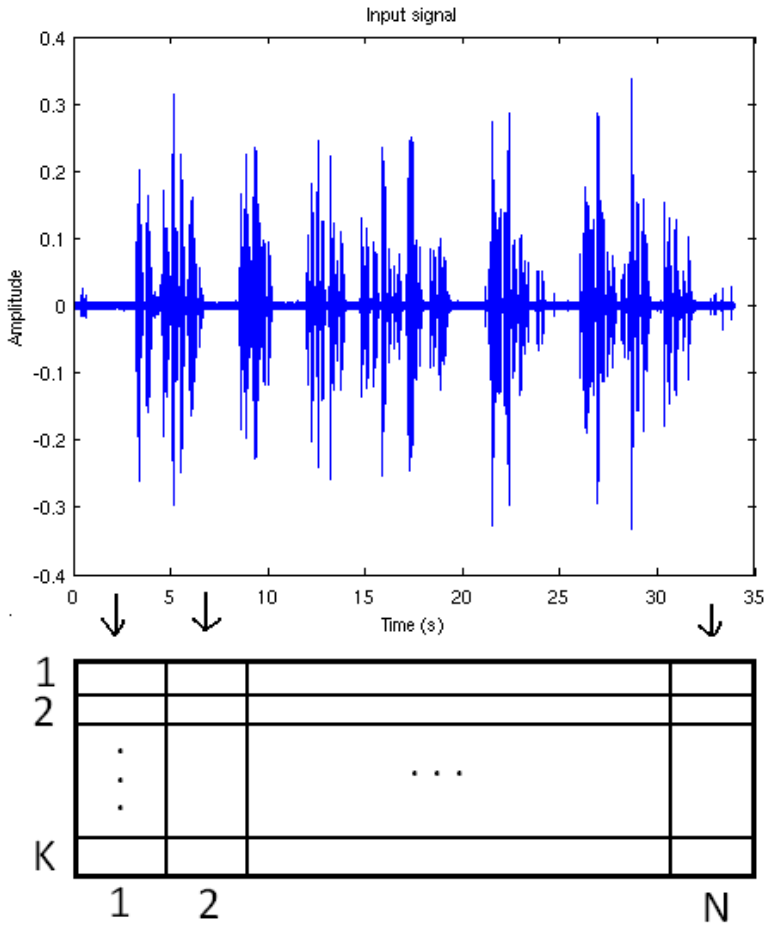

Figure 4. Speech Signal Translated into Acoustic Vectors (AV) using MFCC

The specific design of the MPSOM for Speaker Recognition has 16 clusters arranged in a linear array. The initial radius is 2, reduced to 1 after half of the epochs have been processed, and reduced to 0 after ¾ of the epochs have been reached. The learning rate, α, has an initial value of 0.5 and is linearly decreased on each epoch by an amount equal to α/E, where E is the maximum number of epochs.

For Speaker Identification, the test ***AV***s are compared against each of the trained PSOMs. Only the PSOM of the claimed identity is used for comparison. This comparison is based on minimum Euclidian distance and the statistical parameters stored in the ***P*** matrix. The PSOM with the minimum distance that fulfills the parameters criteria is identified. In this case, the rejection or acceptance of the speaker is done using only the statistical parameters stored in matrix ***P*** and the global distance of the test acoustic vectors to the whole PSOM.

For Speaker Verification, only the PSOM of the claimed identity is used for comparison. In this case, the rejection or acceptance of the speaker is done using only the statistical parameters stored in matrix ***P*** and the global distance of the test acoustic vectors to the whole PSOM. The following section explains in detail the algorithm and rules for rejection or acceptance of a claimed identity.

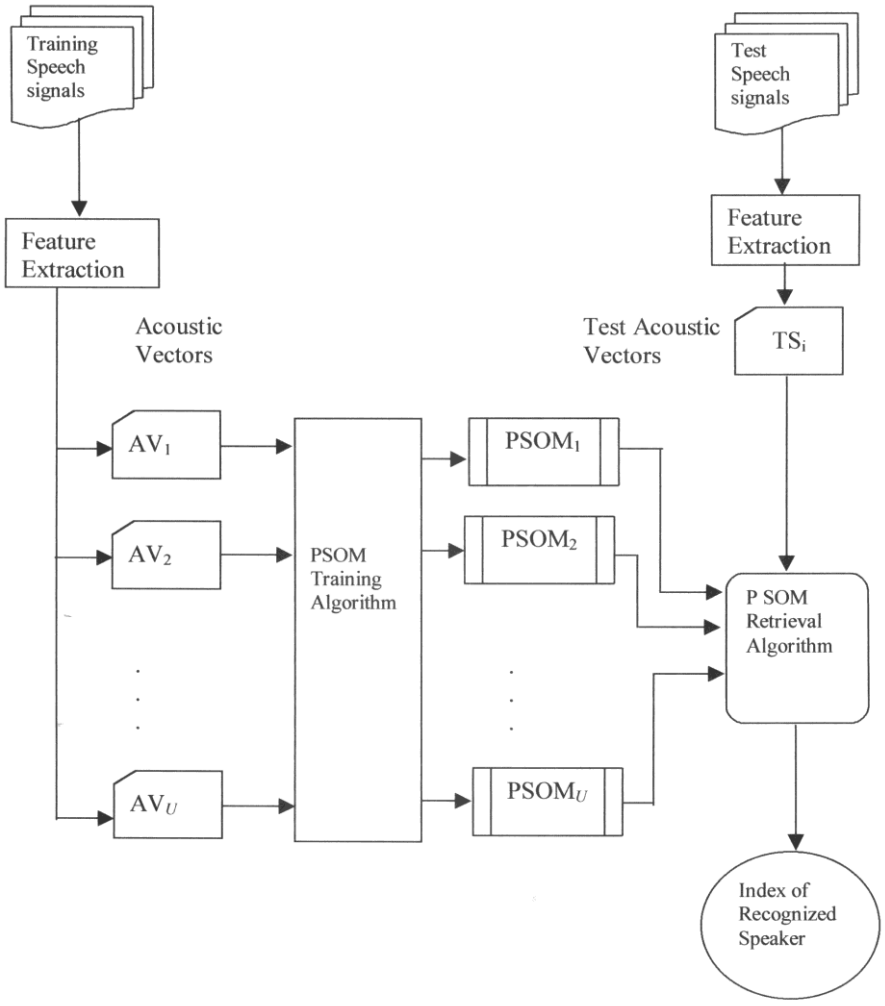

Figure 5. Multiple-PSOM Architecture for Speaker Recognition

6. The Multiple PSOM Retrieval Algorithm

The retrieval algorithm of a MPSOM has 2 versions: one for identification and another one for verification. In this section the identification algorithm is explained since it is more general and in fact contains the verification logic as part of the final decision to identify the closest PSOM found.

In speaker recognition, the acoustic vectors obtained from each user's speech vary from a few hundred to several thousand depending on the length of the speech being processed. The algorithm needs to compute the overall distance between a set of acoustic vectors and the clusters of the MPSOM.

The algorithm is as follows:

Step 1. For all trained users, *u=1 to U*

Step 2. Obtain TS_u a matrix that contains all test acoustic vectors for user *u*, using MFCC. Each row of TS corresponds to a speech signal frame.

Step 3. *minDistance* = ∞
match = 0

Step 4. For all candidate users, *v=1 to U*
d = compute distortion of matrix TS_u against PSOM matrices W_u and P_u (the details are explained in the next section)
if $d \neq \infty$ AND $d <$ *minDistance* then
minDistance = *d*
match = *v* (User *v* has been identified as a candidate)
end if
end for

Step 5. if *match* = 0 then
Speaker *u* is an unknown user (an impostor)
else

$$\mu_v = avg(\mu) = \frac{1}{M}\sum_{j=1}^{M}\mu_{v,j} \quad (1)$$

(Obtain average of μ from PSOM of user *v*)

$$\delta = \left|(minDistance - \mu_v)\right| / \mu_v \quad (2)$$

(Compute variation)

if $\delta > \beta$ then (3) (Speaker *u* is an unknown user)
(an impostor)
else
Speaker *u* matches with trained user *v*
end if
end if

Step 6. End of Algorithm.

The details of computing the overall distortion in step 4 are explained in the next section. For now, it is enough to say that the algorithm returns the average distance of all clustered vectors to the centers of their corresponding clusters. If the number of vectors that fall outside all clusters is greater than 7.5% the algorithm stops and returns the symbolic value

infinity (∞). The value of 7.5% was found experimentally as it is explained in the experiments and results sections. This logic accomplishes two tasks: first, it eliminates impostors in one step, and second, it improves performance since the algorithm stops earlier when the condition is found.

Even with the above validation criterion that filters out most impostors, many times unmatched users have their test acoustic vectors inside the boundaries of the MPSOM of a trained registered user. Step 5 implements the Verification logic. Even if a valid user with a minimum overall distance was identified, this step performs the final validation rule.

Equation 1 is used to calculate the average of parameter μ of the identified candidate MPSOM (μ_v). This average distance is compared to the overall average distance, *d*, obtained by the distortion computation algorithm. Equation 2 is used to compute the ratio of variation between these two average distances. Finally, condition 3 ($\delta > \beta$) is used to reject or accept the claimed identity of the speaker. The value of β, which is the maximum ratio of variation, has been experimentally found to be in the range of 0.25 and 0.30. If a value of 0.25 is used, the ratio of impostor rejection is very high, but the ratio of positive verification decreases. When a value of 0.30 is used, the contrary effect occurs.

7. The Distortion Computation Algorithm

This algorithm calculates the average distance of all acoustic vectors to the centers of all clusters. This is a feature that represents how well the acoustic vectors were clustered by minimizing their distances to the centers of the clusters. This parameter is used to confirm identity because an impostor will have a distortion average larger than the true speaker, even if the impostor's clusters match the true speaker's clusters. The algorithm is as follows:

- For each vector in the training set *TS*, find its Euclidian distance to each cluster *j* in *W*, $j=1, .. M$
- Find the minimum distance to each cluster *j*.
- After all the vectors have been clustered, obtain the average distance to each cluster:
- Obtain the overall distortion by averaging all the non-zero cluster averages

8. Experiment Methodology and Data Selection

To test the algorithms, several known speech corpora were investigated and the CSLU Speech Corpora for Speaker Recognition was selected. The selection was based on the fact that the recordings of these voices are aimed specifically to investigate Speaker Recognition [39].

The data selection consists of 2 sets of 16 speakers. Eight other speakers were also used as impostors for the final experiments, for a total of 40 speakers. The selection objective was to have a balanced population containing equal number of speakers of both genders and spanning a wide range of ages. The age range goes from 16 to over 70 years old. Most papers published by other researchers have used a population from 14 to 30 speakers. Thus a balanced population of 32 trained speakers plus 8 impostors (not trained) is adequate for this research. Obviously, the more speakers evaluated, the more reliable results are expected.

The youngest speaker selected is a boy, who was 16 years old when the first sessions were recorded, and was 18 years old in the last sessions. He was selected for the challenge in recognizing his voice even though it changed notoriously during that period of time.

Experiments Design

One of the goals of this research was to compare the performance of the PSOM against other methods. We chose the vector quantization (VQ) model because this is one of the most widely implemented techniques to identify speakers. Additionally, the experiments were designed to test the performance of Identification, Verification, Impostor Rejection and Text Independence. The experiments were chosen such that we could measure the following metrics:

- Performance of Classification Techniques: VQ vs. Multiple PSOM
- Multiple PSOM Identification Accuracy and Positive Verification
- Multiple PSOM Impostor Rejection using speakers outside the trained population (External Impostor)
- Multiple PSOM Text-Independence using a sentence not part of the training session

It has to be noted that to compare the PSOM and VQ classification techniques, only the identification without verification experiment was done. The reason for this is that the basic algorithm of VQ [16] does not provide a method for verification. The comparison was done on the basis of accuracy, not speed. The reason for this is that the Multiple PSOM method is aimed at high accuracy for applications in Forensic Science.

For the text-independence experiment, CSLU record type "bd" of Session 1 was used for all trained users. The sentence spoken is "Here I am in Miami and Illinois". This sentence does not contain any digits, whereas the PSOMs were trained strictly with sequences of spoken digits.

9. Results and Analysis

The algorithms were implemented using the standard MATLAB® software platform. Several initial experiments were performed to obtain the best method for pre-processing the speech signal. This signal conditional is done in the time domain and consists of several tasks, namely, to eliminate noise, to remove artifacts such as those sounds produced by the lips when they separate to start the first phoneme, to remove periods of silence that do not convey useful information, to normalize the data points and to find the best time frame to segment the signal. After all these signal conditioning tasks the data points were run through a MFCC processor to obtain the Acoustic Vectors. The results of training and testing those vectors are explained in the following sections.

Comparison of Classification Techniques: VQ vs. M-PSOM

To compare the accuracy of both methods, each trained speaker was compared against all the population of trained users of its data set. Four data sets were used: Data Set 1 (16 speakers), Data Set 2 (16 speakers), Data Set 3 is Data Set 1 plus the first 8 speakers of Data Set 2, and Data Set 4 (32 speakers) is the combination of Data Sets 1 and 2. In this way the accuracy is measured with different speaker sets and as a function of data set size. This experiment was done in identification mode only, since the VQ method does not provide a verification algorithm. The results are shown in the following table from which it can be seen that PSOM has a higher accuracy than Vector Quantization as it was expected.

Table 1. Summary of Identification Performance for VQ and PSOM

	Data Set 1 (size=16)	*Data Set 2* (size=16)	*Data Set 3* (size=24)	*Data Set 4* (size=32)	**Average Hit Ratio**
VQ	Hits = 15 Ratio = 93.8%	Hits = 15 Ratio = 93.8%	Hits = 21 Ratio = 87.5%	Hits = 28 Ratio = 87.5%	90.7%
PSOM	Hits = 16 Ratio = 100%	Hits = 15 Ratio = 93.8%	Hits = 22 Ratio = 91.7%	Hits = 29 Ratio = 90.6%	94.0%

Multiple PSOM Identification and Positive Verification Tests

For this part of the experiments, we have to consider the use of the β parameter explained previously. The β parameter defines, along with the size of the cluster, ρ, the verification rule for acceptance or rejection of a claimed identity.

Before the algorithms were tested, special computer programs were run to estimate the number of test acoustic vectors that fall beyond the size of the cluster, ρ, for all trained users, as well as the inter speaker distances in percentage. It was found that a maximum of 7.3% of valid vectors fall beyond the ρ boundary. We set this threshold at 7.5% and use it in the distortion computation algorithm.

The value of β was obtained by averaging the ratio of variation of average distance of the all trained users. It was found that the range $0.25 <= \beta <= 0.30$ is adequate for several applications.

To estimate the accuracy of impostor rejection and positive verification using 16 speakers per data set we have the following possibilities:

- Maximum number of cases = 256 (16 Speakers claiming the identity of 16 speakers)
- Positive Verification Cases = 16
- Maximum Number of Impostor Cases = 256 – 16 = 240

For β=0.3, the number of impostor attempts accepted as valid was 9 therefore:

- Impostor Acceptance Ratio = 9/240 = 0.0375 = 3.75%
- Impostor Rejection Ratio = 100 - Impostor Acceptance Ratio = 96.25%
- Positive Verification Ratio = 16/16 = 100%

For β=0.25, the number of impostor attempts accepted as valid was 4, therefore:

- Impostor Acceptance Ratio = 4/240 = 0.0167= 1.67%
- Impostor Rejection Ratio = 100 - Impostor Acceptance Ratio = 98.33%
- Positive Verification Ratio = 14/16 = 87.5%

These ratios are for impostors within the same population of trained users. The value of β has to be chosen according to the accuracy required for the specific application. For a banking application that requires a password to be spoken on the phone, a β=0.25 would work very well because it is going to reject more than 98% of the impostor attempts. For those valid

users that cannot be verified automatically (12.5%), their calls can be transferred to a customer service representative that will verify the identity in another way.

Multiple PSOM Text-Independence Experiment

The purpose of this experiment is to determine the accuracy of the algorithm when tested with acoustic vectors of valid trained users, using sentences others than those used for training the PSOMs. The training set was built upon several recordings made in sessions 1 and 2 of sentences that contain different sequences of 5 digits.

For this experiment, CSLU record type "*bd*" corresponding to the sentence "Here I am in Miami and Illinois" was used. This sentence is totally different from all sentences used for training. Sixteen test sets corresponding to the sixteen trained users were obtained from these utterances. For this experiment, β=0.275 was chosen.

In this case, the hit ratio is 12/16 or 75.0%. In the previous experiments only one of the six different sentences used for training was used for testing. Now when a totally different sentence is used, the accuracy decreases to 75%, but still shows that the algorithm exhibits text independence.

To explain this phenomenon, let's analyse Table 2 [26] that shows the phonemes contained in the spoken English digits.

The ARPABET code indicates the sequence of phonemes used to utter each word. The American English language has 48 phonemes [1]. Out of these 48 phonemes, the digits use 17, or 35%.

For training purposes, only 35% of the phonemes were used, and still the Multiple PSOM algorithm recognizes 75% of the speakers when a sentence other than the ones used for training was attempted.

To improve the accuracy of the results, the solution is to train the PSOM with sentences that contain more phonemes than just the ones used to pronounce the digits. This is usually the case in normal applications. To prove this hypothesis, the 4 speakers that were rejected were retrained using CSLU record type "*az*" corresponding to the sentence "It's been two years since Dave kept shotguns".

Table 2. Sound Lexicon of English Numbers

Digit	ARPABET Equivalence
Zero	Z-IH-R-OW
One	W-AH-N
Two	T-UW
Three	TH-R-IY
Four	F-OW-R
Five	F-AY-V
Six	S-IH-K-S
Seven	S-EH-V-AX-N
Eight	EY-T
Nine	N-AY-N
Oh	OW

In this case, using β=0.3 the Hit Ratio is 15/16 = 93.75%**,** which proves that to increase the accuracy of the verification algorithm, the speech used for training has to contain the highest number of phonemes practically possible. For some applications the need for text-

independence is not a requirement. Most commercial applications work with spoken numbers only.

Multiple PSOM External Impostor Rejection Test

In these experiments, acoustic vectors from speakers not part of the trained set were used. After the results found in the previous experiment, it was decided to use one of the sentences used in the training session. This way, the system was more challenged.

The experiments consisted in using the acoustic vectors of 8 non-trained users as impostors claiming the identities of the 16 trained (valid) users. Thus, 128 impostor attempts were made in each experiment. For β=0.30 we found that 5 impostor attempts were accepted whereas for β=0.25, only 2 impostor attempts failed to be detected. The next table summarizes the results.

Table 3. Summary of Impostor Rejection Results

β=0.25	Impostor Acceptance = 2/128 = 0.156 = **1.6%** Impostor Rejection = 100% - Impostor Acceptance = **98.4%**
β=0.30	Impostor Acceptance = 5/128 = **3.9%** Impostor Rejection = 100% - Impostor Acceptance = **96.1%**

These results are slightly better than expected, based on the analysis done in the identification and positive verification experiment but clearly show how one can tune a single parameter, β, to arrive at an acceptable precision that is a trade-off between rejecting valid speakers and accepting impostors.

10. Conclusions

A new method for Speaker Recognition, *Multiple Parametric Self-Organizing Maps* (*PSOM*), was designed, implemented and tested successfully yielding an accuracy slightly better than other state-of-the-art methods such as VQ, HMM, GMM and Wavelets. This method is one of the most precise algorithms for Speaker Recognition due to the incorporation of additional parameters to a well known neural network and the extension of the SOM model to multiple dimensions. Other researchers have reported accuracies between 89% and 98%, but mostly below 96%. [8, 32-38].

The high accuracy results obtained with this automated tool are promising for use in many applications, especially for those that require high precision or high impostor rejection ratios such as criminal investigations and forensic science.

The emphasis in the design and implementation of the Multiple Parametric SOM architecture was on precision rather than on speed. However, this same system can perform well in online applications by relaxing the precision of the verification module, that is, by making β=0.3 and by eliminating σ, the standard deviation of the distances, that requires an additional cycle of computations. For customer service applications over the telephone, the accuracy can be as low as 85% because the other 15% that the automated system cannot resolve is transferred to customer service representatives who use additional information to verify the speaker. By simply adjusting a single parameter, β, a company can tune the accuracy of the system and thus the number of people required to resolve the exceptions.

Additional development can be made to improve the performance of this methodology and make it more suitable for real time use. Particularly, the distortion computation algorithm can be run in parallel and not sequentially as it was implemented in this research. This can be accomplished by using several processing engines if the algorithm is implemented in

hardware or multiple threads of execution in a multitasking program if the system is implemented in a personal computer.

References

[1] K. Sri Rama Murty and B. Yegnanarayana, Combining Evidence from Residual Phase MFCC features for Speaker Recognition, IEEE Signal Processing Letters, Vol 13, No. 1, pp. 52-55, January 2006.

[2] Hyoung-Gook Kim, Thomas Sikora, Comparison of MPEG-7 Audio Spectrum Projection Features and MFCC Applied to Speaker Recognition, Sound Classification and Audio Segmentation, IEEE Proceedings on Acoustics, Speech and Signal Processing, Vol. 5, pp. 925-933, May 2004.

[3] Wei-Guo Gong, Li-Ping Yang, Di Chen, Pitch Synchronous Based Feature Extraction for Noise-Robust Speaker Verification,, Vol. 5, pp. 295-298, 2008 Congress on Image and Signal Processing, Vol. 5, 2008

[4] Samuel Kim, Thomas Eriksson, Hong-Goo Kang, Dae Hee Youn, A Pitch Synchronous Feature Extraction Method for Speaker Recognition, IEEE Proceedings on Acoustics, Speech and Signal Processing, pp. 2029-2032, May 2004.

[5] Shung-Yung Lung, Feature extracted from wavelet decomposition using biorthogonal Riesz basis for text-independent speaker recognition, Pattern Recognition, Vol. 41, Issue 10, October 2008, pp. 3068-3070

[6] Shung-Yung Lung, Efficient text independent speaker recognition with wavelet feature selection based multilayered neural network using supervised learning algorithm, Pattern Recognition, Vol. 40, Issue 12, December 2007, pp. 3616-3620

[7] S.-Y.Shung-Yung Lung, Further reduced form of wavelet feature for text independent speaker recognition, Pattern Recognition, Vol. 37, Issue 7, July 2004, pp. 1565-1566

[8] Shung-Yung Lung, Adaptive fuzzy wavelet algorithm for text-independent speaker recognition, Pattern Recognition, Vol. 37, Issue 10, October 2004, pp. 2095-2096

[9] Shung-Yung Lung, Wavelet feature domain adaptive noise reduction using learning algorithm for text-independent speaker recognition, Pattern Recognition, Vol. 40, Issue 9, September 2007, pp. 2603-2606

[10] Bojan Kotnik, Zdravko Kacic, A noise robust feature extraction algorithm using joint wavelet packet subband decomposition and AR modeling of speech signals, Signal Processing, Vol. 87, Issue 6, June 2007, pp. 1202-1223

[11] C. Tantibundhit, J.R. Boston, C.C. Li, J.D. Durrant, S. Shaiman, K. Kovacyk, A. El-Jaroudi, New signal decomposition method based speech enhancement, Signal Processing, Vol. 87, Issue 11, November 2007, pp. 2607-2628

[12] Nengheng Zheng and P.C. Ching, Using Haar transformed vocal source information for automatic speaker recognition, IEEE Proceedings on Acoustics, Speech and Signal Processing, pp. 77-80, May 2004.

[13] Yang Shao, DeLiang Wang, Robust Speaker Recognition Using Binary Time-Frequency Masks, IEEE Proceedings on Acoustics, Speech and Signal Processing, pp. 1589-1592, May 2006

[14] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, A. Stolcke, Modeling prosodic feature sequences for speaker recognition, Speech Communication, Vol. 46, Issues 3-4, Quantitative Prosody Modeling for Natural Speech Description and Generation, July 2005, pp. 455-472

[15] Leena Mary, B. Yegnanarayana, Extraction and representation of prosodic features for language and speaker recognition, Speech Communication, Vol. 50, Issue 10, October 2008, pp. 782-796

[16] William Campbell, Khaled Assaleh and Charles C. Broun, Speaker Recognition With Polynomial Classifiers, IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 4, Part II, pp. 205-212, May 2002.

[17] Ricardo Santana, Rosangela Coelho, Abraham Alcaim, Text-independent speaker recognition based on the Hurst parameter and the multidimensional fractional Brownian motion model, IEEE Transactions on Audio, Speech and Language Processing, Vol 14 No. 3, May 2006.

[18] Afsaneh Asaei, Mohammad Javad Taghizadeh, Marjan Bahrololum, Mohammed Ghanbari, Verified speaker localization utilizing voicing level in split-bands, Signal Processing, Vol. 89, Issue 6, June 2009, pp. 1038-1049

[19] Hong Kook Kim, Seung Ho Choi, Cepstral domain interpretations of line spectral frequencies, Signal Processing, Vol. 88, Issue 3, March 2008, pp. 756-760

[20] Alain Y. Kibangou, Gerard Favier, Blind equalization of nonlinear channels using a tensor decomposition with code/space/time diversities, Signal Processing, Vol. 89, Issue 2, February 2009, pp. 133-143

[21] Y. Linde, A. Buzo & R. Gray, An Algorithm for Vector Quantizer Design, *IEEE Transactions on Communications*, Vol. 28, pp.84-95, 1980.

[22] Ville Hautamaki, Tomi Kinnunen, Pasi Franti, Text-independent speaker recognition using graph matching, Pattern Recognition Letters, Vol. 29, Issue 9, 1 July 2008, pp. 1427-1432

[23] Seiichi Nakagawa, Wei Zhang, Mitsuo Takahashi, Text-independent speaker recognition by combining speaker-specific GMM with speaker adapted syllable-based HMM**,** IEEE Proceedings on Acoustics, Speech and Signal Processing, pp. 81-84, May 2004.

[24] Yih-Ru Wang and Chen-Yu Chiang, A New Common Component GMM-Based Speaker Recognition Method, IEEE Proceedings on Acoustics, Speech and Signal Processing, pp. 770-773, March 2005.

[25] Dalei Wu, Ji Li, Haiqing Wu, α-Gaussian mixture modeling for speaker recognition, Pattern Recognition Letters, Vol. 30, Issue 6, 15 April 2009, pp. 589-594

[26] Longbiao Wang, Norihide Kitaoka, Seiichi Nakagawa, Robust distant speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM, Speech Communication, Vol. 49, Issue 6, June 2007, pp. 501-513

[27] Ismail Shahin, Speaker identification in the shouted environment using Suprasegmental Hidden Markov Models, Signal Processing, Vol. 88, Issue 11, November 2008, pp. 2700-2708

[28] Itshak Lapidot, Hugo Guterman and Arnon Cohen, "*Unsupervised Speaker Recognition Based on Competition Between Self-Organizing Maps*", IEEE Transactions on Neural Networks, Vol. 13, No. 4, pp. 877-887, July 2002.

[29] Lan Wan, Ke Chen and Huisheng Chi, Capture Speaker Information With a Neural Network for Speaker Identification, IEEE Transactions on Neural Networks, Vol. 13, No. 2, pp. 247-252, March 2002.

[30] Laurene Fausett, Fundamentals of Neural Networks. Architectures, Algorithms and Applications, Prentice Hall Inc., 1994. ISBN 0-13-334186-0.

[31] T. Kohonen, The Self-Organizing Map, Proceedings of the IEEE, 78(9): 1464-1480, 1990.

[32] Lawrence Rabiner and Biing-Hwang Juang, Fundamentals of Speech Recognition, Prentice Hall 1993. ISBN 0-13-015157-2.

[33] Tomi Kinnunen, Haizhou Li, An overview of text-independent speaker recognition: From features to supervectors, Speech Communication, Vol. 52, Issue 1, January 2010, pp. 12-40

[34] Gilles Gonon, Frederic Bimbot, Remi Gribonval, Probabilistic scoring using decision trees for fast and scalable speaker recognition, Speech Communication, Vol. 51, Issue 11, November 2009, pp. 1065-1081

[35] Jordi Sole-Casals, Marcos Faundez-Zanuy, Application of the mutual information minimization to speaker recognition and identification improvement, Neurocomputing, Vol. 69, Issues 13-15, August 2006, pp. 1467-1474

[36] Nikos Fakotakis, Anastasios Tsopanoglou, George Kokkinakis, A text-independent speaker recognition system based on vowel spotting, Speech Communication, Vol. 12, Issue 1, March 1993, pp. 57-68

[37] Makoto Yamada, Masashi Sugiyama, Tomoko Matsui, Semi-supervised speaker identification under covariate shift, Signal Processing, Vol. 90, Issue 8, Special Section on Processing and Analysis of High-Dimensional Masses of image and Signal Data, August 2010, pp. 2353-2361

[38] Margarita Kotti, Vassiliki Moschou, Constantine Kotropoulos, Speaker segmentation and clustering, Signal Processing, Vol. 88, Issue 5, May 2008, pp. 1091-1124

[39] Speaker Recognition Corpora from CSLU, Center for Spoken Language Understanding of the Oregon School of Science and Engineering, website: http://cslu.cse.ogi.edu/corpora/spkrec/

Authors

Paul Gomez is the founder and Chief Scientist of Quantum Nanoelectronics, Inc, Florida, U.S.A. Dr. Gomez is a researcher in several areas including Signal Processing and Nanotechnology applied to Medicine. He holds a Bachelor's, Master's and Ph.D. degrees in Computer and Electrical Engineering. Dr. Gomez has been an Adjunct Professor at Florida International University, Devry University and Javeriana University and has been a senior consultant for several Fortune 500 companies for more than 20 years.