Audio Source Classification using Speaker Recognition Techniques

Recognition Technologies Inc., U.S.A.
Technical Report: RTI-20110201-01

Homayoon Beigi[*]

Recognition Technologies, Inc.
3616 Edgehill Road
Yorktown Heights, NY 10598, USA
beigi@recotechnologies.com

Abstract

A practical problem in processing any audio stream is to detect different types of audio and to treat each segment accordingly. This problem may be viewed as a combination of audio segmentation and audio source classification. This paper, treats the latter problem, using a Gaussian Mixture Model (GMM). The problem is formulated as one of identification of several music models and two gender models for speech. First an audio segment is classified as music or speech. Then, the type of musical instrument or the gender of the speaker is tagged. 1400 excerpts of music in different styles from over 70 composers were used together with the speech of 700 male and 700 female speakers. The audio signal was telephone quality sampled at $8kHz$ with μ-law amplitude encoding. A 1% error rate of speech versus music classification and a 1.9% gender classification error rate were achieved at speeds of more than three times real-time on a single core of a multi-core Xeon processor.

1. Introduction

In many practical audio processing systems, it is important to determine the type of audio. For instance, consider a telephone-based system which includes a speech recognizer. Such recognition engines would produce spurious results if they were presented with non-speech, say music. These results may be detrimental to the operation of an automated process. This is also true for speaker identification and verification systems which expect to receive human speech. They may be confused if they are presented with music or other types of audio such as noise. For *text-independent speaker identification* systems, this may result in mis-identifying the audio as a viable choice in the database and resulting in dire consequences!

Similarly, some systems are only interested in processing music. An example is a music search system which would look for a specific music or one resembling the presented segment. These systems may be confused, if presented with human speech, uttered inadvertently, while only music is expected.

The goal of this research is the development of a classification filter which would tag a segment of audio as speech, music, noise, or silence. This problem contains two separate parts. The first part is the segmentation of the audio stream into segments of similar content. This work has been under development for the past few decades with some good results [1, 2, 3].

The second part is the classification of each segment into speech, music, or the rejection of the segment as silence or noise. Furthermore, when the audio type is *human speech*, it is desirable to do a further classification to determine the gender of the individual speaker. Gender classification is helpful in choosing appropriate models for conducting better speech recognition, more accurate speaker verification, and reducing the computation load in large-scale speaker identification.

On the other hand, if the signal of interest is music, it is interesting to be able to determine the specific type of music, for instance in the form of identifying the instrument. Of course, this problem is not quite so simple due to overlap of instruments in orchestral pieces and the sheer number of possible instruments. However, a close approximation to the target instrument and categorization as orchestral or specific types of bands is also useful. We are also interested in an approach which would not require tremendous modeling efforts for every new circumstance which may arise.

To address the instrument identification problem and to be able to cover most types of music, a set of 14 representative instruments or collections of instruments were modeled. We shall see that these models cover an ample space for performing a superb job of classification of *music*, versus *human speech*.

Different approaches with varying perspectives to audio source classification have been reported. One group has tried to identify individual musical instruments [4, 5, 6, 7, 8, 9, 10]; whereas another group has concentrated on classifying speakers based on gender [11, 12, 13, 14, 15, 16]. [17] reports developments in classifying the genre of audio, as stemming from different video sources, containing movies, cartoons, news, etc.

In this research project, a *text* and *language independent* speaker recognition engine is used to achieve these goals by performing audio classification. The classification problem is posed as an identification problem among a series of speech, music, and noise models. Although a very low quality audio, based on highly compressed telephony data, is used, the system achieved a 1% error rate in discriminating between speech and music and a 1.9% error in determining the gender of individual speakers once the audio is tagged as speech.

In Section 2, the prior art in instrument identification and gender classification are discussed. Section 3 describes the audio quality as well as the data collection apparatus. In Section 4, modeling and in Subsection 4.3 the specific GMM-based speaker recognition (used here), have been discussed. Section 5

[*] Homayoon Beigi is the President of Recognition Technologies, Inc. and an Adjunct Professor of Mechanical Engineering at Columbia University

presents the results of these experiments followed by the conclusion in Section 6.

2. Prior Art

In this section, a quick review is given on the two different parts of this research, namely the speech and gender related aspect plus musical instrument identification. Here, different types of audio are modeled using representative samples, designed to cover the space of possible outcomes. It is not our intent to accurately model any one part of this space. Instead, it is desired to have enough coverage such that a rough separation among different types of audio may be achieved.

2.1. Gender Identification

In 1952, Peterson [18] conducted a series of experiments on the 10 common vowels in English. 33 men, 28 women, and 15 children (a total of 76 speakers) were asked to say 10 words (two times each) and their utterances were recorded. The words were designed to examine the 10 vowels in context of an "*h*" to the left and a "*d*" to the right: hid, hɪd, hɛd, hæd, hɑd, hɔd, hʊd, hud, hʌd, and hɝd.

Figure 1 shows the mean value of the fundamental frequency for the vowel in each of the above words, displayed for men, women, and children separately. Note that the fundamental frequency (*formant 0*) does not change much among different vowels. This is the fundamental frequency of the vocal tract based on a normal opening of the vocal folds when one is producing a vowel, but it varies significantly across gender and age. Formants 1 and 2 do vary considerably depending on which vowel is being uttered, however Formant 3 does not [19].

For the past two decades, several different techniques [11, 12, 13, 14, 15, 16], based on the above premise, have been reported for identifying gender. Some effort has also been focused on determining the age groups of individuals based on the above and the concept of *jitter* [19]. [20] proposes using the mean MFCC as an indicator of jitter and states that it is a good indicator of the gender and age of the individual. Here, we are using *Cepstral mean subtraction (CMS)*, and will show great results for gender classification, indicating that gender does not seem to be so correlated to the Cepstral mean. Of course, since age has not been considered in our study, it is possible that the Cepstral mean may still be related to jitter and age.

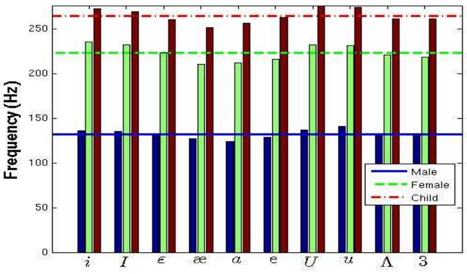

Figure 1: Fundamental Frequencies for Men, Women, and Children while uttering 10 common English vowels – Data From [18]

2.2. Musical Instrument Identification

Martin [4, 21] has used pattern recognition techniques for the problem of musical instrument identification. He uses the *log-lag correlogram* which is adopted from *cochlear models*. This technique is related to the *pitch* which is usually ignored in standard speaker recognition techniques that do not use prosodic features. Since we are not using pitch here, it is fundamentally different from our approach. For a robust and universal resolution, the objective is to determine *timbre* and not be dependent on values related to *pitch* and *sonority* (see [19] for motivation).

[7] also uses cepstral coefficients for conducting musical instrument recognition. However, it uses very complex features which are connected to the dynamics of musical pieces and maps the frequencies to the Bark frequecy scale [22] which is similar to the Mel-Frequency mapping in that it is also based on the *psychophysical power law of hearing* [19]. However, the complex set of features as well as heuristics make this approach too impractical for the purpose of a simple and universal prefilter. The approach of [7] is more suitable for accurate recognition of instruments and is inherently much more costly.

As previously discussed, every effort has been made to make a simple and flexible model which may be easily modified in order to be able to handle a finer granularity of audio types. Modeling very specific aspects, although attractive, may result in too much complexity and reduction of practicality.

In the following section, the experimental apparatus is described, aimed at showing the effectiveness of a GMM-based speaker recognition system in determining timbre as well as gender. Great effort has been expended to ensure that practicality is not sacrificed. For instance, it would have been simple to take on clean audio at high sampling rates to do this demonstration. However, the real word is seldom so giving. Therefore, audio data with low sampling rates has been chosen.

3. Apparatus

The speech part of the apparatus was described in some detail in [23], therefore, only a brief summary is given here. The speech data was collected using μ-Law amplitude encoding [24] at a sampling rate of $8kHz$. The audio was then immediately converted to $8kHz$ *high-efficiency advanced audio coding* format (**HE-AAC**) [25] which is a very aggressive, lossy, and low-bit-rate audio compression technique. **HE-AAC** was used to stream the audio to a server through *flash*. In turn, the audio was converted back to μ-Law and subsequently converted to 16-bit $8kHz$ linear Pulse Code Modulation (**LPCM**), inside the classification engine. The original signal was recorded using 100 stations with different hardware, at random. 700 male and 700 female speakers were selected, completely at random, from over $70,000$ speakers in our database. The speakers were non-native speakers of English, at a variety of proficiency levels, speaking freely. This introduced significantly higher number of pauses in each recording, as well as more than average number of humming sounds while the candidates would think about their speech. The segments were live responses of these non-native speakers to test questions in English, aimed at evaluating their linguistic proficiency.

An equal amount of music was chosen to create a balance in the quantity of data, reducing any bias toward speech or music. The music was downsampled from its original quality to $8kHz$, using 8-bit μ-Law amplitude encoding, in order to match the quality of speech. The 1400 segments of music were chosen at random from European style classical music, as well as jazz, Persian classical, Chinese classical, folk, and instructional performances. Most of the music samples were orchestral pieces with some solos and duets present.

4. Audio Classification Modeling

The following two subsections briefly describe the models. Each identification model only occupies about 72 kB of storage and includes the statistics for representing Gaussian mixtures. In this case, instead of modeling the voice of an individual, specific speech and music models were built to perform the classification task.

4.1. Speech and Gender Modeling

According to the discussions of Section 2.1, a simple gender recognition system would be possible by estimating the fundamental frequency or determining the location of the formants for different vowels. However, here we would like to steer away from hard-coded algorithms which may work for a portion of the population, but fail miserably for some outliers. An effort has been made to pool together a diverse set of male and female speakers (see Table 1) in order to create a male and a female model. This method has the flexibility of being able to adapt [23] to new outliers introduced in the future.

Model No.	Category	Model Description	Enrollment Length (s)	No. of Samples
1	Noise	Noise	120	–
2	Speech	Female	3315	57
3	Speech	Male	7536	122
4	Music	Accordion	138	–
5	Music	Bassoon	126	–
6	Music	Clarinet	135	–
7	Music	Clavier	109	–
8	Music	Gamelon	155	–
9	Music	Guzheng	121	–
10	Music	Guitar	174	–
11	Music	Oboe	110	–
12	Music	Orchestra	213	–
13	Music	Piano	141	–
14	Music	Pipa	203	–
15	Music	Tar	146	–
16	Music	Throat	76	–
17	Music	Violin	222	–

Table 1: Audio Models used for Classification

4.2. Music Modeling

Much in the same spirit as described in Section 4.1, an effort has been made to choose a variety of different instruments or sets of instruments to be able to cover most types of music. Table 1 shows these choices. A total of 14 different music models were trained to represent all music. A conscious effort was made to pick these instruments in such as way that they would cover different types of timbres [21].

4.3. A Gaussian Mixture Model Recognizer

The RecoMadeEasy[1] speaker recognition engine was used. This engine is a **GMM**-based *text-independent* and *language-independent* engine. It uses models for the speaker and the competing models to conduct an open-set identification task. The population in the identification task was 17, as shown in Table 1. The models are parameters for collections of multivariate normal density functions which describe the distribution of the Mel-Cepstral features [19] for speakers' enrollment data. This distribution is represented by Equation 1.

$$p(\mathbf{x}) \quad = \quad \frac{1}{(2\pi)^{\frac{d}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}} exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{1}$$

where $\mathbf{x}$, $\boldsymbol{\mu} \in \mathscr{R}^d$ and $\mathbf{\Sigma} : \mathscr{R}^d \mapsto \mathscr{R}^d$.

In Equation 1, $\boldsymbol{\mu}$ is the mean vector where,

$$\boldsymbol{\mu} \overset{\Delta}{=} \mathscr{E}\{\mathbf{x}\} \overset{\Delta}{=} \int_{-\infty}^{\infty} \mathbf{x}\, p(\mathbf{x}) d\mathbf{x} \tag{2}$$

The *sample mean* approximation for Equation 2 is,

$$\boldsymbol{\mu} \approx \frac{1}{N}\sum_{i=0}^{N-1} \mathbf{x}_i \tag{3}$$

where N is the number of samples and $\mathbf{x}_i$ are the MFCC [19].

The *Covariance* matrix is defined as,

$$\mathbf{\Sigma} \overset{\Delta}{=} \mathscr{E}\left\{ (\mathbf{x}-\mathscr{E}\{\mathbf{x}\})\, (\mathbf{x}-\mathscr{E}\{\mathbf{x}\})^T \right\} = \mathscr{E}\left\{ \mathbf{x}\mathbf{x}^T \right\} - \boldsymbol{\mu}\boldsymbol{\mu}^T \tag{4}$$

The diagonal elements of $\mathbf{\Sigma}$ are the variances of the individual dimensions of $\mathbf{x}$. The off-diagonal elements are the covariances across the different dimensions.

The *unbiased estimate* of $\mathbf{\Sigma}$, $\tilde{\mathbf{\Sigma}}$ is given by the following,

$$\tilde{\mathbf{\Sigma}} = \frac{1}{N-1}\left[\mathbf{S}_{xx} - N(\boldsymbol{\mu}\boldsymbol{\mu}^T) \right] \tag{5}$$

where the *sample mean* $\boldsymbol{\mu}$ is given by Equation 3 and the *second order sum matrix*, $\mathbf{S}_{xx}$ is given by,

$$\mathbf{S}_{xx} \overset{\Delta}{=} \sum_{i=0}^{N-1} \mathbf{x}_i \mathbf{x}_i{}^T \tag{6}$$

The features used by the recognizer are *Mel-Frequency Cepstral Coefficients* (MFCC). Unfortunately, due to the shortage of space, not much detail may be presented in this section. Reference [19] describes details of such a **GMM**-based recognizer.

5. Results

Figure 2 shows the classification error between speech and music. It has been computed for different segments of audio from 2.5s to 60s long. As the graph shows, the ideal amount of audio needed for classification is close to 10s. This produces nearly the same results as with 60s of data, but it is much more practical. Indeed, even reducing to 2.5s does not degrade the results by much. Given the poor audio quality, this is quite promising, especially as far as the speech is concerned. In fact, most of the errors come from segments which were actually speech and were classified as music. A major contributor to this is that many of the speech recordings have a loud babbling sound in the background, since they have been recorded in a public test area with close to a hundred people taking tests in cubicles.

Figure 3 shows the *conditional error rate* of gender classification. This is the percentage of the speech segments which were correctly identified as speech, but were tagged with the wrong gender. In this case, we can see that the the accuracy of the gender classification is a bit more tied to the amount of data. However, still a 10s segment seems to be quite practical for producing decent results.

As the length of audio was reduced, certain segments were rejected for not containing enough audio. As expected, the number of rejections was at its maximum of 16, when the length

[1] RecoMadeEasy® is a trademark of *Recognition Technologies*

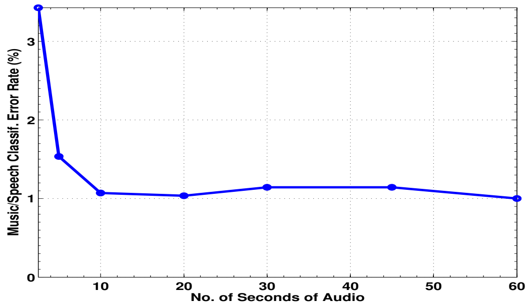

Figure 2: Speech vs. Music Classification Error (2800 samples – 700 male speakers, 700 female speakers, and 1400 music segments) suggests an optimal segment length of about 10 seconds for maximum performance with the least amount of audio

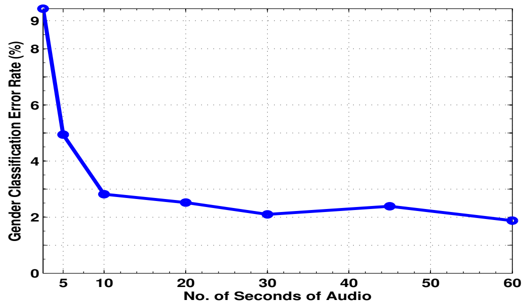

Figure 3: Gender Classification Error (relative to samples which are tagged as speech). Figure suggests an optimal segment length of about 10 seconds to achieve a practical sample length with an acceptable error.

of audio was reduced to 2.5s. Some examples of music which were rejected at 2.5 were the *Lullaby* by *Aram Khachaturian*, *Bolero* by *Maurice Ravel*, *Eine Kleine Nachtmusik* by *Mozart*, and *Jesu, Joy of Man's Desiring* by *Bach*. These are some of the better known examples of the 16 rejected pieces. They are all very soft and slow. Therefore, chances of getting silences or very low energy segments in a 2.5 second segment were quite high. Still this is only a 0.57% of the total. Due to their nature, these should not be considered as errors, since in fact the 2.5 segments did not match speech or music; they were indeed silence and were correctly classified.

The identification process was faster than three times real-time, on a single core of a 2.8GHz *Intel Xeon* processor. In addition, due to the simple formulation, the process was easily parallelized. Using two quad-core processors, more than 24 times real-time performance was achieved in obtaining these results, making it ideal for pre-filtering audio signals prior to introducing them to systems such as speech recognizers or music understanding modules.

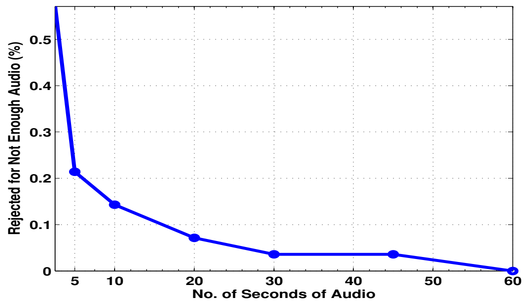

Figure 4: Percent of Audio Files Rejected for Not Having Enough Audio

6. Conclusion

A practical filter has been designed using speaker identification to detect speech and music, and to further classify speech into one uttered by different genders and music coming from different sources. In fact, looking at the details of the results from the musical categories, it is often the case that a solo instrument is recognized properly with the second and third choice being similar instruments. For example, a Guitar and a Tar often follow as alternate choices for each other. The closest musical instrument to the male voice seems to be the Bassoon, and sometime Oboe and clarinet followed the female voice.

In most cases where speech was detected as music, the second or third choice seemed to be speech. Also, in many of these cases, the speaker was thinking out loud, by humming. This happened quite often in these samples, since all the speakers were non-native speakers and were responding to live questions, requiring a long thought process.

Although the classification system produced the type of instrument based on Table 1, quantifying the accuracy in these cases is very complex, due to the many instances of orchestral pieces in the test. Such quantification and evaluation is being considered and will be presented in future publications. Depending on the results, improvements will be proposed in optimizing the list of instrument models used for classification.

In this study, the quality of the audio was chosen to be quite challenging in order to keep the system within practical means. It is expected that by increasing the number of instrument models and the sampling frequency, much more detailed instrument recognition would be achievable. Future work is on the way for classifying the signature (timbre) of specific instruments within a class of instruments. This is especially interesting, in order to be able to determine the authenticity of vintage and high quality instruments from fakes. Presently, work on enrolling *Tars* from different instrument manufacturers is being pursued, in order to determine the accuracy of recognizing an unseen instrument. A master instrument maker can often hear a few notes and determine the manufacturer of vintage instruments.

7. References

[1] Stephane H. Maes Homayoon S. M. Beigi, "Speaker, channel and environment change detection," 1997, IBM Research Tech. Rep. RC21022, Yorktown Heights, NY.

[2] Homayoon S.M. Beigi and Stephane S. Maes, "Speaker, channel and environment change detection," in *Proc. of the World Congress on Automation*, May 1998.

[3] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environemnt and channel change detection and clustering via the bayesian inromation criterion," in *IBM Res. Tech. Rep.*, 1998, Yorktown Heights.

[4] K. D. Martin and Y. E. Kim, "Musical instrument identification: A pattern-recognition approach," in *136th Meeting of the Acoustical Society of America*, Oct 1998.

[5] A. A. Livshin and X. Rodet, "Musical instrument identification in continuous recordings," in *Proc. of the 7th Int. Conf. on Digital Audio Effects*, Oct 2004, pp. 1–5.

[6] G. Mazarakis, P. Tzevelekos, and G. Kouroupetroglou, "Musical instrument recognition and classification using time encoded signal processing and fast artificial neural networks," in *Adv. in Artificial Intell.*, vol. 3955 of *Lect. Notes in Comp. Sci.*, pp. 246–255. Springer, Berlin, 2006.

[7] A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," in *ICASSP*, Jun 2000, vol. 2, pp. 753–756.

[8] A. Eronen, "Comparison of features for musical instrument recognition," in *IEEE Workshop on the Apps. of Signal Proc. to Audio and Acoustics*, Oct 2001, pp. 19–22.

[9] E. Benetos, M. Kotti, and C. Kotropoulos, "Large scale musical instrument identification," in *Proc. of Sound and Music Computing Conf.*, Jul 2007, pp. 283–286.

[10] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrogram: A new musical instrument recognition technique without using onset detection nor f_0 estimation," in *ICASSP*, 2006, vol. 5, pp. 229–232.

[11] D.G. Childers, Ke Wu, K.S. Bae, and D.M. Hicks, “Automatic recognition of gender by voice,” in *ICASSP*, Apr 1988, vol. 1, pp. 603–606.

[12] Ajmera J., “Effect of age and gender on lp smoothed spectral envelope,” in *The IEEE Odyssey*, Jun 2006, pp. 1–4.

[13] A. S. Naini and M. M. Homayounpour, “Speaker age interval and sex identification based on jitters, shimmers and mean mfcc using supervised and unsupervised discriminative classification methods,” in *The 8th Int. Conf. on Sig. Proc.*, 2006, vol. 1.

[14] E. Scheme, E. Castillo-Guerra, K. Englehart, and A. Kizhanatham, “Practical considerations for real-time implementation of speech-based gender detection,” in *Proc. of the Iberoamerican Cong. in Patt. Reco.*, Nov 2006.

[15] F. Yingle, Y. Li, and T. Qinye, “Speaker gender identification based on combining linear and nonlinear features,” in *Proc. of the WCICA*, Jun 2008, pp. 6745–6749.

[16] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Nöth, “Age and gender recognition for telephone applications based on gmm supervectors and support vector machines,” in *ICASSP*, Apr 2008, pp. 1605–1608.

[17] P. Dhanalakshmi, S. Palanivel, and V. Ramalingam, “Classification of audio signals using aann and gmm,” *Applied Soft Computing*, vol. 11, no. 1, pp. 716 – 723, 2011.

[18] G. Peterson and H. L. Barney, “Control methods used in a study of the vowels,” *JASA*, vol. 24, no. 2, pp. 175–185, Mar 1952.

[19] Homayoon Beigi, *Fundamentals of Speaker Recognition*, Springer, New York, 2011, ISBN: 978-0-387-77591-3.

[20] Naini A. S. and M. M. Homayounpour, “Speaker age interval and sex identification based on jitters, shimmers and mean mfcc using supervised and unsupervised discriminative classification methods,” in *The 8th Int. Conf. on Sig. Proc.*, 2006, vol. 1.

[21] K. D. Martin, *Sound-Source Recognition: A Theory and Computational Model*, MIT, MA, 1999, PhD Thesis.

[22] E. Zwicker, G. Flottorp, and Stanley Smith Stevens, “Critical band width in loudness summation,” *JASA*, vol. 29, no. 5, pp. 548–557, 1957.

[23] Homayoon Beigi, “Effects of time lapse on speaker recognition results,” in *IEEE Conf. on DSP*, Jul 2009, pp. 1–6.

[24] ITU-T, “G.711 Pulse Code Modulation (PCM) of Voice Frequencies,” ITU-T Recommendation, Nov. 1988.

[25] S. Meltzer and G. Moser, “Mpeg-4 he-aac v2 – audio coding for today’s digital media world,” 2005, http://www.ebu.ch/fr/technical/trev/trev_305-moser.pdf.

Homayoon Beigi, holds three research positions. For the past few years, he has conducted research and development in the fields of Biometrics, Pattern Recognition and Internet-Commerce. As the President of Recognition Technologies, Inc., he is conducting research toward the production of a series of Speaker Recognition, Language Modeling and Signature Recognition line of RecoMadeEasy$^{(TM)}$ software engines. His work as the Vice President of Internet Server Connections, Inc. includes the development of the multiple-award winning Commerce Made Easy$^{(R)}$ software, an elaborate Electronic Commerce system used world-wide. In addition as an Adjunct Professor of Mechanical Engineering, he has taught "Applied Signal Recognition and Classification", "Speech and Handwriting Recognition" and "Digital Control" at the mechanical engineering and electrical engineering departments of Columbia University. He was a Research Staff Member at the IBM T.J. Watson Research Center from 1991 to 2001 where he conducted research on Speaker Recognition, Language Modeling, Aggressive Search Techniques, Speech Recognition, On-Line Handwriting Recognition, Software Architecture, Control Theory and Neural Network Learning.