Open-Ended Design and Performance Evaluation of a Biometric Speaker Identification System

Ravi P. Ramachandran, Robi Polikar and Kevin D. Dahm
Rowan University
ravi@rowan.edu, polikar@rowan.edu, dahm@rowan.edu

Sachin S. Shetty
Tennessee State University
sshetty@tnstate.edu

***Abstract*— It is very important that biometrics education, particularly at the undergraduate level, keeps pace with the rapidly growing global market. This paper describes a senior level project in speech biometrics that fits in a variety of courses in order to reach out to many students. The project has broad learning outcomes, namely, enhanced application of math skills, software implementation skills, interest in biometrics, ability to carry out open-ended design and communication skills. Assessment results based on the analysis of the success of the students (refereed publications and enrolling in graduate programs), student surveys related to the learning outcomes and a target versus control group survey show that the project was successful.**

I. Introduction

The teaching of design through a project based experience has been shown to not only increase open-ended design skills but also further reinforce concepts in basic engineering and mathematics through vertical integration [1][2][3]. Vertical integration is the principle of having a project or experiment in an upper level course that builds upon concepts gained through experiments and/or projects performed in a lower level course. Students will realize that the courses are part of a flow that contributes to a unified knowledge base. Moreover, a project based on a modern topic that is highly relevant to today's marketplace increases student interest [4].

Biometrics is a modern topic with primary applications in the commercial, government and law enforcement sectors. It is highly significant in enhancing cybersecurity which is of key global importance. Also, the biometrics market continues to grow very rapidly and is expected to reach $7.1 billion by 2012 with a compound annual growth rate of 21.3 percent [5]. There is much research interest in different types of biometric systems notably speaker identification [6][7][8]. Speaker identification systems have advantages including ease of use and implementation, low cost and high user acceptance [6]. In addition, they can be easily integrated (no special hardware required) with many devices including desktops, laptops, cell phones, wireless access points, iPhones, iPads and PDAs.

There is an acute need for biometrics education at the undergraduate and graduate levels. Many institutions world-wide have an established graduate program in biometrics and offer senior level undergraduate elective courses [9][10] in the area. The University of West Virginia offers a Bachelor of Science in Biometric Systems. The U.S. Naval Academy has a Biometrics Research Laboratory with an aim to enhance undergraduate biometrics education [10] where a senior undergraduate elective course on Biometric Signal Processing is offered that integrates lecture and laboratory experiences. Configuring a new undergraduate program and/or a new biometrics laboratory requires enormous resources that are beyond the reach of most institutions even in the best of economic times. This paper describes one senior level project in an NSF sponsored effort to vertically integrate biometrics across an existing undergraduate curriculum. At the senior undergraduate level, the biometrics projects are designed to fit in a variety of courses in order to reach out to many students, The projects have broad learning outcomes.

This paper gives the project details, learning outcomes and assessment results of the design and implementation of a biometric speaker identification system. The project has the attributes of teaching open-ended design and software implementation of a complete system, reinforcing basic math and engineering concepts, achieving vertical integration, focusing on a modern topic relevant to today's society and using a real-life speech database. This project can be assigned in a variety of senior level signal processing courses like pattern recognition, speech processing and biometric systems.

II. Learning Outcomes

In implementing a speaker identification system, students go through each step, namely, preprocessing (voice activity detection and preemphasis), feature extraction, classification (training and use in rendering a decision) and performance evaluation. The KING database is used to show students that robustness to mismatched training and testing conditions is a significant practical issue. The open-ended aspects include researching different robust features, implementing different classifiers and investigating feature and classifier fusion to augment performance. The student learning outcomes of the project include:

1) Enhanced application of math skills.
2) Enhanced software implementation skills.
3) Enhanced interest in biometrics.
4) Enhanced ability to read research papers and apply algorithms (like robust feature extraction) to achieve a better design thereby providing research experience.
5) Enhanced communication skills.
6) Comprehension of the importance of vertical integration [1]][3] in that students realize that their experiences are part of a curricular flow.

III. Description of Project

The speech database along with the training and performance evaluation phases are first described. Then, the actual project assignment is discussed. Students are taught the mathematical background and concepts of linear prediction, feature extraction, vector quantizer (VQ) design and the decision logic used in speaker identification.

A. KING Database

The KING corpus was created for research in the area of speaker identification [11]. It was collected at two locations, namely, San Diego (26 speakers) and New Jersey (25 speakers). There are ten sessions for each speaker (numbered 01 to 10). Sessions were recorded a week to a month apart. The speech was passed through a standard telephone handset, transmitted through a local telephone exchange to a long distance service and back to the local exchange,

978-1-4673-0219-7/12/$31.00 ©2012 IEEE

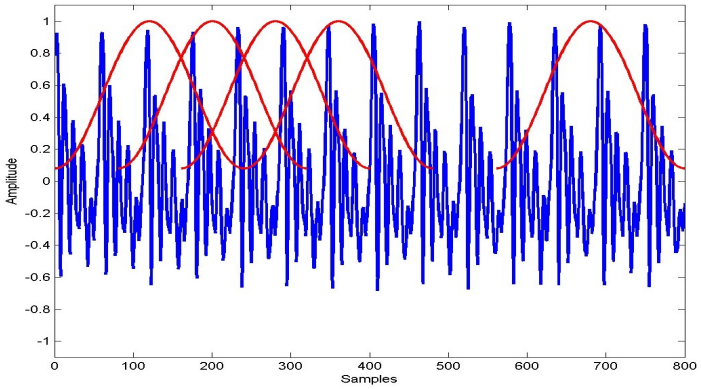

Fig. 1. Frame by frame processing

and then recorded from an analog telephone patch. There is both narrow-band communication channel and noise distortion. The speech is sampled at 8 kHz.

In this project, it is compulsory to use the 26 San Diego speakers. Investigating the New Jersey speakers is one of the optional tasks. A peculiar anomaly of the narrow-band San Diego data is the phenomenon known as "The Great Divide". There is an apparent change in the spectral characteristics of the narrow-band channel between sessions 1-5 and sessions 6-10. This involves a difference in spectral slope for the composite transfer functions in the two sets. Speaker identification algorithms generally perform well within the divide and perform poorly across the divide as a result. It is a challenge to get a high performance across the divide.

The database has ten directories labeled s01 to s10. The directory indicates the session number. Each directory has 26 speech files in ASCII format labeled spkr1.dat to spkr26.dat. There is one speech file for each of the 26 speakers (for example spkr15.dat is the speech file for speaker 15).

B. Preprocessing of the Speech

Students are exposed to frame-by-frame processing of a speech signal as depicted in Figure 1. This is common to every speech utterance that is processed. The length of each frame is 30 ms. The overlap between consecutive frames is 20 ms. Each frame is multiplied by a Hamming window and effectively represents the middle 80 samples of its entire 240 sample length.

For each frame, a voice activity detector is used to discriminate between speech-like high energy segments and silence [12]. Students can choose to implement the method in [12] or formulate their own algorithm. The speech is preemphasized by the filter $1 - 0.95z^{-1}$ and for each speech-like frame, the autocorrelation method of linear prediction is used to get a 12th order polynomial $A(z)$.

An additional step is to identify and use only speech-like frames with well defined formant frequencies [11]. The procedure is to find the roots of $A(z)$ and count the number of roots that (1) have an imaginary part greater than 0, (2) a magnitude greater than or equal to 0.88 and (3) an angle between a frequency of 300 Hz and 3700 Hz. A frame is selected for feature extraction if the number of roots that satisfy the above criteria is greater than or equal to 3. This is known as linear prediction based frame selection.

C. Feature Extraction

For each selected frame, calculate the 12 dimensional linear predictive cepstrum (CEP), adaptive component weighted (ACW) cepstrum

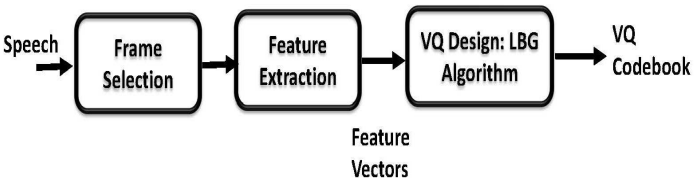

Fig. 2. Vector quantizer classifier training

[11], postfilter (PFL) cepstrum [11], pole filtered mean removed cepstrum (PFMRCEP) [13], mean removed ACW cepstrum (MRACW), pole filtered mean removed ACW cepstrum (PFMRACW) [14] and mean removed PFL cepstrum (MRPFL). Seven different feature vectors of dimension 12 are computed. The motivation of using these features is (1) because six of the seven (except CEP) enhance robustness to channel effects, (2) to do a performance comparison and (3) to investigate fusion strategies. With seven features, there are effectively seven speaker identification systems that are configured.

D. System Training

In many real-life applications, a limited amount of training data is available. Each experiment is performed by training the system on one of the ten sessions. Each session has only one utterance for each speaker. Hence, for a particular speaker, one speech utterance is used for training the vector quantizer (VQ) codebook as shown in Figure 2. A VQ classifier, consisting of 26 codebooks (one for each speaker), is designed for each of the seven features using the Linde-Buzo-Gray (LBG) algorithm. The distortion measure is the squared Euclidean distance.

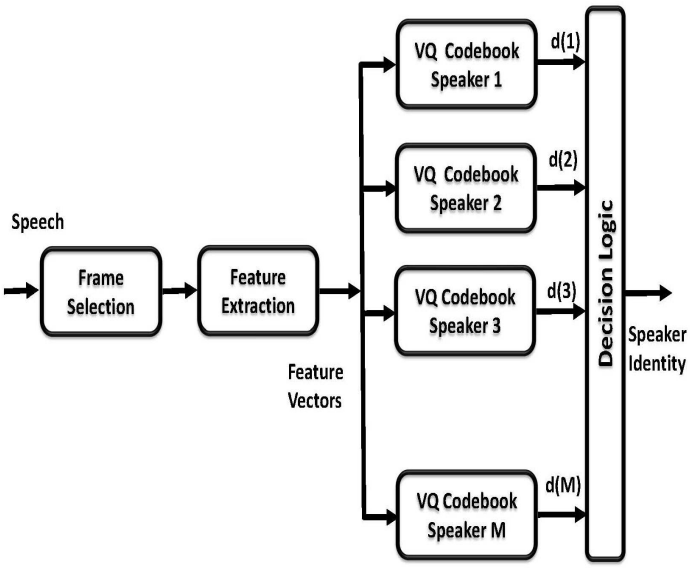

Fig. 3. Vector quantizer classifier for speaker identification

E. Performance Evaluation

The VQ system for processing a test speech utterance and identifying a speaker is shown in Figure 3. A test utterance from one of the speakers is converted to a set of test feature vectors after frame selection. Consider a test feature vector. This is quantized by each of the VQ codebooks. The quantized vector is that which is closest to the test feature vector in terms of the squared Euclidean distance. There

are $M = 26$ different distances recorded, one for each codebook. This process is repeated for every test feature vector. The distances are accumulated over the entire set of feature vectors such that $d(i)$ is the accumulated distance for codebook i. The codebook that renders the smallest accumulated distance identifies the speaker. When many utterances are tested, the identification success rate (ISR) is the number of utterances for which the speaker is identified correctly divided by the total number of utterances tested.

F. Project Assignment

The deliverables for the project are a formal report and MATLAB code organized in a modular fashion with a brief description on how to run the code. Specific project tasks include:

1) Listen to the speech files. Are there any perceivable differences "across the divide"?
2) Read the tutorial papers [6][12] and write a critical synopsis on biometric speaker recognition. The paper [7] was not available but will be assigned the next time the project is run.
3) Train on sessions s01 to s05, one at a time. Use VQ codebooks of size 64. Record the ISR for the nine remaining test sessions for each of the seven features. Do the speaker identification experiments with and without linear prediction based frame selection.
4) Use hypothesis testing and confidence interval estimation to see if certain features achieve a statistically better performance. Use this statistical approach to see if linear prediction based frame selection improves performance.

Some suggestions were made for open-ended design with the objective of augmenting the ISR.

1) Research other robust features.
2) Use other classifiers like Support Vector Machines, Neural Networks and Gaussian Mixture Models. Perform classifier fusion.
3) For a given classifier (VQ or other), examine feature fusion strategies. Examples are decision level fusion, probability level fusion and Borda count.
4) Combine feature and classifier fusion.

IV. Assessment Results

A. Open-ended Design

What did students do for open-ended design? Most students followed the suggestions given. The most commonly implemented idea was a fusion strategy. A few students tried the mel-frequency cepstrum (MFCC) feature [15] without and with mean removal (MRMFCC). Some ancillary experiments were conducted like attempting different VQ codebook sizes and processing the New Jersey recordings of the KING database. The most impressive work was by two students who independently experimented with the MFCC and tried Soong-Rosenberg fusion (which was not suggested) [15] and probability level fusion. In Soong-Rosenberg fusion, a weighted linear combination of the VQ codebook distances of different features is used for deciding the speaker identity. The weights are determined from the training data. In probability fusion, the VQ distances are coverted to probabilities to achieve normalization in the range 0 to 1. The maximum value of the linear combination of the probabilities for different features identifies the speaker. This work resulted in a conference paper [16] for which the main results for the four best features are given in Table I. The results in Table I are for the case when the classifier is trained on session s01. Fusion improves performance. The two students are now in graduate school working in the area of pattern recognition.

Another positive outcome is regarding one student who, in parallel with taking the course, did a year-long project on speaker identification in the presence of G.729 coding distortion. This also resulted in a paper [17] and the student is starting graduate school in the area of biometric speech processing.

B. Survey of Learning Outcomes and Vertical Integration

A survey relating to the learning outcomes was given to the 15 students participating in the project. Table II gives the results. There was no response less than 3 (Neutral) for any of the questions. Student perception of vertical integration was assessed by giving them a list of sophomore and junior/senior courses and asking them whether the material learned in these courses had a connection with the speaker identification project. The sophomore courses included Basic Circuits, Electronics, Digital Circuits and Mathematics (differential equations, linear algebra, complex variables and probability and statistics). The junior/senior level courses included Digital Signal Processing, Communications, Advanced Electronics, Control Systems, Microprocessors and Computer Architecture. All students selected Digital Signal Processing, 80 percent picked Mathematics and 53.3 percent picked Control Systems.

C. Target Group Versus Control Group

In order to obtain a quantitative analysis whose significance can be statistically evaluated, a target group of 15 students that participated in the biometrics project is compared with a control group of 14 students that did not participate. The survey is designed such that its true intent, to determine whether students awareness and interest in biometrics increased, will be hidden. This concept was applied in [18] to evaluate interest in biomedical engineering. In the survey, the students are asked four questions on the areas of electrical / computer engineering that they find interesting, would take an elective course in, consider as a career option and consider as a graduate school thesis topic. In each question, there will be many options, in which biometric related answers will be randomly distributed. For example, the question on elective courses lists 22 courses of which four biometrics related courses are buried. For each of the 22 courses, the student selects a score of 0 (no interest), 1 (somewhat interested) or 2 (very interested). A biometric interest factor for this question is calculated based on what scores the students select for the four biometrics related courses only. For this question, the maximum interest factor is 8. The same principle is used for the other questions to calculate a total biometric interest factor and a normalized biometric interest factor in the range 0 to 1. This normalized biometric interest factor is found for both the target and control groups. The mean and standard deviation for the target group is 0.57 and 0.17 respectively. For the control group, the mean and standard deviation is 0.45 and 0.20 respectively. The relatively high standard deviations is due to the small population of the two groups. Future studies will have larger student populations. Also, two students in the control group had a very high interest factor which means that they have an interest in biometrics even though they did not do the project. A one-tailed t-test with unequal variances indicates that the target group has a higher biometric interest factor with a p-value of 0.047. This means that the difference is statistically significant with an 95.3% confidence. This result gives a 95.3% confidence that the difference between the two groups is not due to chance, but in fact due to the target group being exposed to the project.

V. Summary and Conclusions

The biometric speaker identification project has achieved many learning outcomes, given the students a perception of the usefulness

Identification Success Rates (%)									
Features/Fusion	Session 2	Session 3	Session 4	Session 5	Session 6	Session 7	Session 8	Session 9	Session 10
Soong-Rosenberg Fusion	84.6	73.1	80.8	84.6	57.7	46.2	53.9	53.9	50.0
Probability Level Fusion	84.6	73.1	80.8	80.8	53.9	42.3	42.3	46.2	53.9
MRMFCC	73.1	69.2	76.9	73.1	42.3	26.9	50.0	57.7	42.3
PFMRACW	76.9	57.6	73.1	73.1	42.3	34.6	46.2	30.8	46.2
PFMRCEP	76.9	69.2	76.9	80.8	38.5	23.1	11.5	26.9	30.8
MRACW	69.2	38.4	46.2	50.0	23.1	11.5	19.2	23.1	23.1

TABLE I

PERFORMANCE ON THE KING DATABASE WITH TRAINING ON SESSION S01 (TAKEN FROM [16])

1 - Strongly disagree, 2 - Disagree, 3 - Neutral, 4 - Agree, 5 - Strongly Agree			
Statement	Mean	Median	Standard Deviation
The project helped reinforce MATLAB software skills.	4.27	4	0.70
The project enriched mathematical and analytical skills.	4.13	4	0.55
The project helped reinforce written communication skills.	4.00	4	0.65
The project provided background in pattern recognition and biometrics as it applied to speech processing.	4.47	5	0.64
The project helped gain valuable experience in open-ended design/research on speech based biometric systems.	4.40	4	0.51
I am now more likely to follow popular media news / developments / programs that relate to biometrics as compared to before doing the project.	3.87	4	0.74

TABLE II

PROJECT OUTCOME SURVEY RESULTS

of vertical integration and stimulated interest in biometrics. Students implemented a complete system, did a performance evaluation and understood one of the main problems in biometrics research, namely the robustness due to mismatched training and testing conditions. In addition, three students have published their work.

VI. ACKNOWLEDGEMENT

This work was supported by the National Science Foundation through Grants DUE-1122296 and DUE-1122344.

VII. REFERENCES

1) C. Trullemans, L. De Vroey, S. Sobieski and F. Labrique, "From KCL to class D amplifier", *IEEE Circuits and Systems Magazine*, pp. 63–74, First Quarter, 2009.
2) P. Jansson, R. P. Ramachandran, J. L. Schmalzel and S. A. Mandayam, "Creating an agile ECE learning environment through engineering clinics", *IEEE Transactions on Education*, Vol. 53, No. 3, pp. 455–462, August 2010.
3) Y. Tang, L. M. Head, R. P. Ramachandran and L. M. Chatman, "Vertical integration of system-on-chip concepts in the digital design curriculum", *IEEE Transactions on Education*, Vol. 54, No. 2, pp. 188–196, May 2011.
4) Y. Tsividis, "Turning students on to circuits", *IEEE Circuits and Systems Magazine*, pp. 58–63, First Quarter, 2009.
5) The Global Biometrics Market, BCC Research, January 2007. (http://www.sbwire.com/news/view/18766).
6) A. K. Jain, A. Ross and S. Prabhakar, "An introduction to biometric recognition", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 14, No. 1, January 2004.
7) R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues", *IEEE Circuits and Systems Magazine*, pp. 23–61, June 2011.
8) J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J.-F. Bonastre and D. Matrouf, "Forensic speaker recognition", *IEEE Signal Processing Magazine*, pp. 95–103, March 2009.
9) S. Cotter, "Laboratory Exercises for an Undergraduate Biometric Signal Processing Course", *ASEE Annual Conference*, Louisville, Kentucky, June 2010.
10) R. W. Ives, Y. Du, D. M. Etter and T. B. Welch, "A Multidisciplinary Approach to Biometrics", *IEEE Transactions on Education*, Vol. 48, No. 3, pp. 462–471, August 2005.
11) M. S. Zilovic, R. P. Ramachandran and R. J. Mammone, "Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions", *IEEE Trans. on Speech and Audio Processing*, Vol. 6, No. 3, pp. 260–267, May 1998.
12) T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors", *Speech Communication*, Vol. 52, pp. 12–40, 2010.
13) R. P. Ramachandran and K. R. Farrell, "Fast pole filtering for speaker recognition", *IEEE Int. Symp. on Circuits and Systems*, Geneva, Switzerland, pp. V-49–V-52, May 2000.
14) A. L. Swanson, R. P. Ramachandran and S. H. Chin, "Fast adaptive component weighted cepstrum pole filtering for speaker identification", *IEEE Int. Symp. on Circuits and Systems*, Vancouver, Canada, pp. V-612–V-615, May 2004.
15) T. F. Quatieri, *Discrete Time Speech Signal Processing Principles and Practice* Prentice Hall PTR, 2002.
16) G. Ditzler, J. Ethridge, R. P. Ramachandran and R. Polikar, "Fusion Methods for Boosting Performance of Speaker Identification Systems", *IEEE Asia Pacific Conf. on Circuits and Systems*, Kuala Lampur, Malaysia, December 6–9, 2010.
17) R. Mudrowsky, R. P. Ramachandran and S. S. Shetty, "The Affine Transform and Feature Fusion for Robust Speaker Identification in the Presence of Speech Coding Distortion", *IEEE Asia Pacific Conf. on Circuits and Systems*, Kuala Lampur, Malaysia, December 6–9, 2010.
18) R. Polikar, R. P. Ramachandran, L. M. Head and M. Tahamont, "Introducing Multidisciplinary Novel Content Through Laboratory Exercises on Real World Applications", *ASEE Annual Conference*, Honolulu, Hawaii, June 2007.