DOI: 10.22413/ijastems/2026/v12/i6/5

Quantitative Evaluation of Zero Frequency Filtering for Speaker Identification Using Hybrid MFCC-GCI Features and SVM Classification

J T Pramod, S. D. Nafisa Fathima, S. Divya, K. Gowthami
Dept. of ECE, Aditya College of Engineering, Madanapalle, AP, India

Abstract— This paper presents a speaker classification framework that integrates Glottal Closure Instant (GCI) features extracted using Zero Frequency Filtering (ZFF) along with Mel Frequency Cepstral Coefficients (MFCCs). Both feature vectors, vocal tract spectral properties and glottal excitation features are combined to provide complementary speaker-discriminative information. To perform speaker classification, Support Vector Machine (SVM) classifier with Error-Correcting Output Codes (ECOC) is used.A speech corpus from Kaggle containing speeches of five prominent leaders namelyBenjamin Netanyahu, Jens Stoltenberg, Julia Gillard, MargaretTacher and Nelson Mandela with 7500 utterances sampled at 16 kHz is used. Experimental results show the overall classification accuracy of 97.56%, GCI Identification Rate of 97.23%, Miss Rate of 2.77%, False Alarm Rate of 5.67%, and Mean Absolute Timing Error of 1.238 ms.The per-speaker accuracy is above 95% consistently for all speakers, with one speaker Nelson Mandelaachieving perfect classification. The experimental results confirm that fusing source-level epoch features with spectral coefficients meaningfully strengthen speaker classification performance.

Index Terms— Speaker Identification, Zero Frequency Filtering, Glottal Closure Instants, MFCC, Support Vector Machine, Epoch extraction.

I. INTRODUCTION

Speaker identificationdetermines the identity of a personfrom the features hidden deep in their speech utterances. Automatic speech recognitiondecodes spoken information whereas the speaker identification gives information about who is speaking, irrespective of the utterances themselves. This unique property of the speaker identification system makes it valuable for access control, forensic sciences, and human-computer interaction.

The human speech works as a unique biometric signal which is modelled by three interacting subsystems: the respiratory system, which supplies the airflow with continuously varying pressure; the phonatory system consisting of vibrating vocal folds generate quasi-periodic pulses which act as a glottal source; and the articulatory system consisting of resonant cavities enforce a speaker-specific spectral envelope on the source signal. The source-filter model consisting of the excitation source and vocal tract system provides a foundation for extracting both spectral and excitation-based features suitable for recognition [11].

Mel Frequency Cepstral Coefficients (MFCCs) are widely used in speaker recognition systems because they effectively capture the spectral characteristics of speech that are closely related to human auditory perception [7]. These features mainly provide important speaker-specific information as they represent the resonant properties of the vocal tract. However, MFCCs do not capture the excitation or glottal source information generated by the vocal folds during speech production.

Glottal Closure Instants (GCIs), which occur when the vocal folds close during each pitch cycle, contain useful excitation-related information such as pitch variations and excitation strength. These features contribute source information along with MFCC based spectral features and improve speaker discrimination. Among various GCI detection techniques, the Zero Frequency Filtering (ZFF) method proposed by Murty and Yegnanarayana is preferred because of its simplicity, computational efficiency, and robustness [1]. In this work, both MFCC and GCI features are combined to form a speaker identification framework to improve speaker identification performance.

II. LITERATURE SURVEY

The speaker identification has undergone several important developments over the years. Early systems were based on simple template matching. They were gradually replaced by Gaussian Mixture Model (GMM) approaches. These approaches were capable of modelling the statistical distribution of speech features for individual speakers [10]. The introduction of the GMM–Universal Background Model (GMM-UBM) framework further improved system robustness. It adapted a general speaker-independent background model to enrolled speakers. These developments significantly enhanced the reliability and accuracy of speaker identification systems.

Research on epoch or Glottal Closure Instant (GCI) extraction has also received considerable attention. Murty and Yegnanarayana introduced the Zero Frequency Filtering (ZFF) technique as a robust and computationally simple method for epoch detection that does not rely heavily on linear prediction analysis [1]. Alku discussed various glottal inverse filtering methods for estimating voice source information [2], while Gerratt, Kreiman, and Garellek highlighted the additional complexity involved in analyzing continuous speech signals compared to sustained phonation [3]. Srinivas et al. then implemented the ZFF technique on FPGA hardware for real-time applications [4].

DOI: 10.22413/ijastems/2026/v12/i6/5

Gurugubelli and Yuvapalla addressed the numerical stability issues associated with cascaded integration in ZFF systems [5]. Reddy et al. proposed a wavelet-based GCI detection approach with improved robustness in noisy conditions [6].

Support Vector Machines (SVMs) were later introduced into speaker verification systems by Wan and Renals.They demonstrated that SVM-based classifiers could achieve performance comparable to GMM-based systems [8]. The theoretical basis of SVMs, particularly their margin maximization capability, is well established in the literature [9]. Deep learning approaches experimented recently such as x-vector and ECAPA-TDNN systems have achieved state-of-the-art performance in speaker recognition tasks. These methods, however require large datasets and high computational resources, making them less suitable for low resource environments. Detailed quantitative evaluation of the ZFF method specifically for speaker identification within a complete speaker identification framework is still limited even though significant research has been carried out in both speaker recognition and GCI detection. The present work attempts to address this research gap.

.

III. THEORETICAL BACKGROUND

A. Source-Filter Model

The speech signal $s(t)$ is decomposed as the convolution of a glottal source $g(t)$, a vocal tract impulse response v(t), and a lip radiation term $l(t)$ as,

$$s(t) = g(t) * v(t) * l(t) \quad (1)$$

In the frequency domain the same can be represented as [11]

$$S(f) = G(f) \cdot V(f) \cdot L(f) \quad (2)$$

MFCC features primarily characterize the vocal tract filter $V(f)$, while GCI features characterize the source $G(f)$. Jointly exploiting both the glottal source and the vocal tract filter response provides a complete representation of speaker identity.

B. MFCC Extraction

MFCC extraction proceeds through six stages. A pre-emphasis filter,

$$H(z) = 1 - \alpha z^{-1} \, (where, \alpha \approx 0.97) \quad (3)$$

compensates for spectral roll-off. The signal is then divided into overlapping frames of 25 ms with a 10 ms shift; each frame is multiplied by a Hamming window to reduce spectral leakage. The resultant windowed signal is converted frequency domain by 512-point FFT which yields the power spectrum.This power spectrum is then filtered by 26 triangular mel-scale filters spanning 80 Hz to 8 kHz. The logarithm of each filter energy is computed, and a Discrete Cosine Transform is applied to the resultant which produces 13 cepstral coefficients [7]. Delta and delta-delta derivatives are also considered to expand the static vector to 39 dimensions which captures the temporal trajectory of spectral features.

C. Zero Frequency Filtering

ZFF detects the instants where the glottis closes producing impulse like excitations which dominate the zero-frequency neighbourhood of the speech spectrum [1]. The algorithm consists of four steps: (1) first-order derivatives to suppress DC and promote transient events; (2) cascaded double integration to implement the zero-frequency resonator

$$y_2[n] = 2y_2[n-1] - y_2[n-2] + y_1[n] \quad (4)$$

(3) moving-average trend removal over a window of approximately 1–2 pitch periods; and (4) detection of positive zero-crossings or local maxima in the resultant detrended signal as Glottal Closure Instants or Epochs. The resultant features extracted include the pitch period T_0, excitation strength at each GCI, and epoch rate.

D. SVM with ECOC

Support Vector Machines (SVMs) classify data by identifying an optimal decision boundary, known as the hyperplane $w^T x + b = 0$, which maximizes the margin between different classes. In real-world speech data, perfect separation between classes is rarely possible. To handle such cases, slack variables ξ_i and a regularization parameter C are introduced to maintain a balance between maximizing the margin and minimizing classification errors. The binary SVM framework is extended using the Error Correcting Output Code (ECOC) approach as the speaker identification is a multi-class problem involving a number of speakers. In this method, the multi-class problem is divided into multiple binary classification tasks, and the final class decision is obtained by combining the outputs using minimum Hamming distance decoding. Radial Basis Function (RBF) kernel is employed to improve the classification performance for complex and non-linear speech patterns, The RBF kernel is mathematically represented as:

$$K(x_i, x_j) = \exp\left(-\gamma \left||x_i - x_j\right|\right|^2) \quad (5)$$

This kernel transforms the original feature vectors into a higher-dimensional space where different speakers can be better separated [9]. Hence, the SVM classifier becomes highly effective for distinguishing speakers based on both MFCC and GCI feature representations.

IV. PROPOSED SYSTEM ARCHITECTURE

The block diagram of the proposed system shows the sequential flow of processing from raw speech input to the final speaker identification output. The proposed system processes raw speech through five sequential modules: preprocessing, epoch extraction, MFCC extraction, feature fusion and classification and performance evaluation.

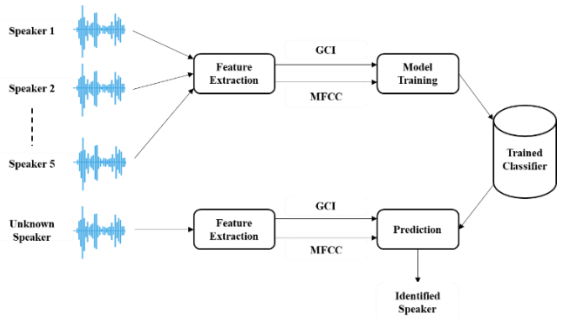

Figure I. Block Diagram of Speaker Identification System

DOI: 10.22413/ijastems/2026/v12/i6/5

A. Preprocessing

The recorded speech signals are first converted into mono format by averaging the stereo channels. The amplitude of the signal is then normalized to a unit peak using,

$$x_{norm}(n) = x(n)/max|x(n)| \quad (6)$$

This normalization ensures uniform signal amplitude across all recordings. To reduce the effect of background noise, a noise profile is estimated from silence or non-speech recordings and used for spectral noise reduction. The DC offset present in the signal is removed by subtracting the mean value of the waveform. In addition, a Butterworth high-pass filter with a cut-off frequency of 80 Hz is applied to suppress low-frequency noise and electrical hum components.

After preprocessing, voiced and unvoiced regions of speech are separated using short-time energy, zero-crossing rate, and autocorrelation-based pitch detection techniques. Since Glottal Closure Instants (GCIs) occur only in voiced speech regions, only the voiced frames are forwarded to the epoch extraction stage. This preprocessing pipeline improves the quality and reliability of both spectral and excitation-based feature extraction used in the speaker identification system.

B. Feature Fusion

For each utterance segment, a 39-dimensional MFCC vector (13 static + 13 delta + 13 delta-delta) is computed at the frame level and averaged across voiced frames. GCI features such as mean pitch period, standard deviation of pitch period, mean excitation strength, and epoch rate are combined to form a unified feature vector $F = [MFCC; GCI]$. Feature vectors are z-score normalized prior to SVM training using statistics computed on the samples used for training.

C. Training and Testing Protocol

The speech database is divided speaker-wise using a file-based splitting approach. 70% of the speech files are used for training and the remaining 30% are used for testing. This procedure makes sure that the same speech utterance does not appear in both the training and testing sets.This prevents data leakage and also enables a fair evaluation of system performance. The training phase uses labelled hybrid feature vectors consisting of both MFCC and GCI-based features to build the speaker identification model.

An SVM classifier based on the Error Correcting Output Code (ECOC) is trained using these extracted feature vectors. During testing, unknown speech samples are processed in the same manner and their corresponding feature vectors are provided to the trained model for prediction. The final speaker decision is achieved using the minimum-distance ECOC decoding rule. This rule combines the outputs of multiple binary classifiers to identify the speaker class with high likelihood.

V. EXPERIMENTAL SETUP

The speech database used in this work were obtained from Kaggle. It consists of speeches of five prominent leaders namely Benjamin Netanyahu, Jens Stoltenberg, Julia Gillard, MargaretTacher and Nelson Mandela. All speech samples were made available with a sampling frequency of 16 kHz with 16-bit PCM encoding which preserves sufficient speech quality for analysis. For each speaker, the number of speech utterances available were 1,500 with durations ranging from 3 to 5 seconds, resulting in a total of 7,500 speech recordings for five speakers. Out of these, 5,250 audio files were used for training and the remaining 2,250 audio files were reserved for testing the speaker identification process.The speech recordings were not completely noise-free, realistic acoustic variations were present in the recordings in all five speaker recordings in the dataset. All stages of the speaker identification i.e., preprocessing, feature extraction, GCI detection, training, and testing are implemented using MATLAB. The speaker identification system employs an SVM classifier with a Radial Basis Function (RBF) kernel.In this, important parameters such as the kernel parameter γ and regularization parameter C are optimized through cross-validation using the training dataset to achieve improved classification performance.

VI. RESULT AND DISCUSSION

A. Speaker Identification Performance

Table I summarizes the overall system performance. The system provides an overall accuracy of 97.56%.This accuracy demonstrates that the hybrid MFCC-GCI feature set, combined with SVM-ECOC classification, achieves robust speaker discrimination across all five speakers. The detection rate matches accuracy which confirms consistent correct prediction rates across the corpus. A false alarm rate of 2.44% is obtained which indicates that minimum inter-speaker confusion is evident. The execution time is 77.12swhich is good enough for offline batch processing.For real-time deployment, further optimization may be required.

Table I. Speaker Identification Performance

Parameter	Value
Overall Accuracy	97.56%
Detection Rate	97.56%
False Alarm Rate	2.44%
Execution Time	77.12 seconds

B. Per-Speaker Accuracy

Table II shows the speaker-wise identification accuracy obtained by the proposed system. All five speakers achieve an accuracy greater than 95%, indicating that the classifier performs consistently well across different speakers without showing significant bias toward any particular individual. Among the speakers, Speaker 5 (Nelson Mandela) achieves 100% accuracy, which suggests that the speech characteristics of this speaker are highly distinctive and easily separable from the others. In contrast, Speaker 2(Jens Stoltenberg) records the lowest accuracy of 95.78%,

DOI: 10.22413/ijastems/2026/v12/i6/5

which may be due to similarities in vocal characteristics such as pitch or spectral patterns with other speakers in the dataset.

Table II. Per-Speaker Identification Accuracy

S No	Speaker	Accuracy (%)
1	Benjamin Netanyahu	97.33
2	Jens Stoltenberg	95.78
3	Julia Gillard	97.78
4	Margaret Tacher	96.89
5	Nelson Mandela	100.00

Overall, the variation in speaker-wise accuracy is very small, demonstrating that the proposed speaker identification model is well balanced and capable of generalizing effectively across multiple speakers. These results indicate that the combination of MFCC and GCI-based features, along with the SVM-ECOC classifier, provides reliable and consistent speaker discrimination performance.

C. Confusion Matrix Analysis

Table III shows the confusion matrix indicating the strong performance of the proposed system.The diagonal elements being dominantindicates that most test samples are correctly classified. Speaker 1 is correctly identified in 438 out of 450 test samples, with only a few samples being misclassified as Speakers 2 and 4. Speaker 2 shows comparatively higher misclassification than the other speakers, where a small number of samples are incorrectly recognized as Speakers 1 and 3. This could be due to similarities in vocal characteristics such as pitch or spectral patterns between these speakers.

Speaker 3 achieves 441 correct identifications, while Speaker 4 records 436 correct predictions with only minor confusion involving Speakers 2 and 3. Speaker 5 achieves perfect classification accuracy, with all 450 test samples correctly identified and no misclassifications observed. This indicates that Speaker 5 (Nelson Mandela) possesses highly distinctive speech characteristics that are easily separable from the other speakers. Finally, the confusion matrix indicates that the proposed MFCC-GCI based speaker identification system performs reliably across different speakers. The minor confusion observed among Speakers 1 to 4 suggests slight overlap in their vocal characteristics.

Table III. Confusion Matrix

True Class \ Predicted Class	1	2	3	4	5
1	438	7	2	3	0
2	7	431	9	3	0
3	9	0	441	1	0
4	1	6	7	436	0
5	0	0	0	0	450

D. Waveform with GCI Locations

The waveform plot provides a visual comparison between the speech signal, the reference GCI locations, and the GCIs detected by the proposed algorithm. From the graph it is seen that the detected GCIs indicated by red markers closely follow the reference GCIs indicated by green markers throughout the signal duration. The detected points are distributed evenly across the waveform, this indicates that the algorithm is capable of identifying glottal closure instants over the entire speech segment rather than only in isolated regions.

The strong alignment between the reference GCIs and detected GCIs supports the strong GCI detection performance obtained numerically. The high Identification Rate (IDR) and low Miss Rate (MR) indicate that most of the true glottal events are successfully detected. In addition, the low Mean Absolute Timing Error (MATE) confirms that the detected GCIs are temporally accurate, with only small deviations from the reference locations. The overall plot demonstrates that the proposed GCI detection method achieves reliable and accurate performance even though there are few additional detections present which contribute to the False Alarm Rate (FAR).

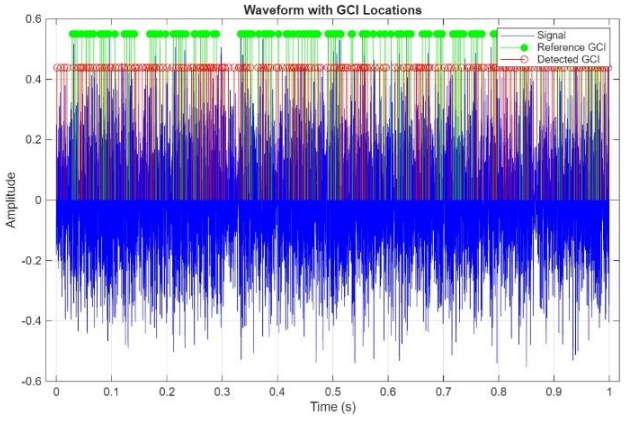

Figure II. Waveform with GCI Locations

E. Histogram of GCI Timing Errors

The histogram of GCI timing errors shows how closely the detected GCIs match the reference GCI locations in terms of timing. It can be observed that most of the timing errors are concentrated within a very small range of around 1.2ms. This narrow distribution of the histogram indicates that the variation in timing error is low. It also shows that the detected GCIs are consistently aligned with the reference points.

This distribution confirms the good temporal accuracy of the proposed GCI detection method. The concentration of errors near zero indicates that the algorithm is capable of detecting GCIs with high timing precision and only minor

DOI: 10.22413/ijastems/2026/v12/i6/5

deviations from the reference GCIs. Overall, the histogram demonstrates that the proposed method provides reliable time mapping of GCIs across the speech signal.

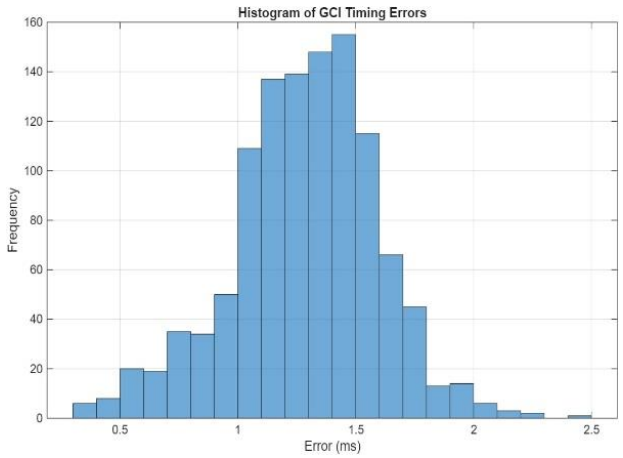

Figure III. Histogram of GCI Timing Errors

F. GCI Detection Metrics

Table IV presents the performance of the proposed GCI detection method.The performance is evaluated using a pseudo ground truth generated by combining the outputs of three well-known GCI detection algorithms which are DYPSA, SEDREAMS, and YAGA. The obtained Identification Rate (IDR) of 97.23% shows that the proposed ZFF-based detector is capable of identifying almost all the reference GCIs accurately. The low Miss Rate (MR) of 2.77% indicates that only a small number of GCIs are missed during detection. The False Alarm Rate (FAR) of 5.67% confirms that the number of incorrect detections is also minimal.The Mean Absolute Timing Error (MATE) of 1.238ms indicates that the detected GCIs are closely aligned with the reference positions, showing good temporal precision.

This observation is further supported by the timing error histogram. The timing errors are concentrated well within a narrow range around 1 to 1.5ms. Overall, these results indicate that the proposed GCI detection approach provides reliable detection accuracy along with precise temporal localization of glottal events. Hence, this technique is suitable for speech and speaker recognition applications.

Table IV. GCI Detection Performance

Metric	Value
IDR	97.23%
MR	2.77%
FAR	5.67%
MATE	1.238 ms

G. Discussion

The speaker identification system proposed is of hybrid natures which achieves an overall accuracy of 97.56%. This demonstrates the effectiveness of combining spectral features i.e., MFCCs with excitation-based GCI features. The fusion representation of features provides complementary information about both the vocal tract and the excitation source which improves the overall discriminative capability of the system. In addition to this, the SVM-ECOC classifier effectively handles the multi-speaker classification problem and performs well in the high-dimensional hybrid feature space.

The waveform visualization further shows that the detected GCIs closely align with the actual glottal closure locations throughout the speech signal. The alignment accuracy between detected and reference GCIs remains high, which explains the low Mean Absolute Timing Error (MATE) even though the detected GCIs are comparatively fewer than the dense reference GCIs. This indicates that the ZFF-based GCI extraction method provides precise time mapping of GCIs. The strong GCI detection performance achieved in this work can also be attributed to the voiced/unvoiced segmentation stage, which helps reduce false detections in unvoiced speech regions and improves the overall robustness of the system.

VII. CONCLUSION & FUTURE SCOPE

A speaker identification system which is designed using the combination of ZFF-based epoch features with MFCC features using an SVM-ECOC classifier has been evaluated in this work [1,12]. The proposed system achieves an overall speaker identification accuracy of 97.56% across five speakers, while the individual speaker accuracies remain consistently above 95%. In addition to this, the GCI detection module achieves a high Identification Rate (IDR) of 97.23%, a low Miss Rate (MR) of 2.77%, and a Mean Absolute Timing Error (MATE) of 1.238ms. These results indicate that the ZFF algorithm is capable of providing both reliable GCI detection and accurate time mapping. The overall performance confirms that combining spectral information from the vocal tract with excitation-based GCIs can significantly improve speaker identification performance.

The outcomes of this work also pave the way for several possibilities of future research. More advanced deep learning-based GCI detection methods, such as Convolutional Neural Networks (CNNs) or Transformer-based architectures which are trained directly on raw speech signals can be experimented on. These advance models may further improve detection accuracy and robustness when hybrid temporal and spectral features are utilized. The proposed system can also be extended to larger and more diverse multilingual speech databases, particularly involving Indian languages and different speaking styles. Real-time implementation of the proposed framework on embedded or FPGA-based hardware platforms can be investigated for practical applications as a future experiment. Future work may also explore multimodal approaches that combine speech with visual or physiological signals to improve system robustness in noisy and challenging environments.

REFERENCE

DOI: 10.22413/ijastems/2026/v12/i6/5

[1] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," IEEE Trans. Audio, Speech, Lang. Process., vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[2] P. Alku, "Glottal inverse filtering analysis of human voice production: A review of estimation and parameterization methods of the glottal excitation and their applications," Sadhana, vol. 36, no. 5, pp. 623–650, Oct. 2011.

[3] B. R. Gerratt, J. Kreiman, and M. Garellek, "Comparing measures of voice quality from sustained phonation and continuous speech," J. Speech Lang. Hear. Res., vol. 59, no. 5, pp. 994–1001, Oct. 2016.

[4] T. Srinivas, B. Pradhan, and A. Kumar, "FPGA implementation of zero frequency filter for real-time epoch extraction," in Proc. Int. Conf. Signal Process. Commun. (ICSC), 2018.

[5] K. Gurugubelli and A. K. Yuvapalla, "Stable implementation of zero frequency filtering of speech signals," in Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP), 2019, pp. 6730–6734.

[6] A. Reddy, K. S. R. Murty, and B. Yegnanarayana, "Epoch extraction using wavelet-based methods for speech signals," IEEE Signal Process. Lett., vol. 27, pp. 910–914, 2020.

[7] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust. Speech Signal Process., vol. 28, no. 4, pp. 357–366, Aug. 1980.

[8] V. Wan and J. Renals, "Speaker verification using sequence discriminant support vector machines," IEEE Trans. Speech Audio Process., vol. 13, no. 2, pp. 203–210, Mar. 2005.

[9] V. Vapnik, The Nature of Statistical Learning Theory. New York, NY, USA: Springer, 1995.

[10] L. Rabiner and B. H. Juang, Fundamentals of Speech Recognition. Englewood Cliffs, NJ, USA: Prentice Hall, 1993.

[11] B. Yegnanarayana, "Extraction of speaker-specific information from speech signals," IEEE Trans. Speech Audio Process., vol. 5, no. 1, pp. 1–12, Jan. 1997.

[12] H. Murty and B. Yegnanarayana, "Epoch-based features for speaker recognition," IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 5, pp. 1549–1557, Sep. 2006.