International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 11 (2015) pp. 28389-28399

http://www.ripublication.com

Speaker Identification Using Relevance Vector Machine

C. Sunitha
Head of Department, Department of CA & SS, Sri Krishna Arts and Science College, Affiliated to Bharathiar University, Coimbatore, India.
phdsunithascholar@gmail.com
Dr. E. Chandra
Professor, Department of Computer Science, Bharathiar University, Coimbatore, India. crcspeech@gmail.com

Abstract

Speech being a unique characteristic of an individual is widely used in speaker verification and speaker identification. Speaker verification and identification are used in many applications. Dual Tree- Complex Wavelet Transform (DT-CWT) is used to extract features from speech signal. The Complex Wavelet Transform is a tool that uses a dual tree of wavelet filters to find the real and imaginary parts of complex wavelet coefficients. The advantage of DT-CWT is more efficient and less redundant. The speaker verification is done by Relevance Vector Machine (RVM).Experimental result gives better Performance in terms of FRR, FAR and execution time.

Keywords: Dual Tree- Complex Wavelet Transform, Relevance Vector Machine, Speaker Recognition.

Introduction

Speaker recognition can be defined as the task of establishing the identity of speakers from their voices. The ability of recognizing voices of those familiar to us is a vital part of oral communication between humans. Research has considered automatic computer based speaker recognition since the early 1970's taking advantage of advances in the related field of speech recognition.

Speaker recognition has two major applications: speaker identification and speaker verification. Identification of unknown individual speaker from many templates is defined as speaker identification. Verification is one where individual speaker's voice is matching with another one in a template. Speaker recognition is efficient tool and it is used by both the government sector and industrial purpose. For example, the Australian Government organization Centre link uses speaker verification for the authentication of Welfare recipients using telephone transactions [1].

Potential applications of speaker recognition include forensics [2], access security, phone banking, web services [3], personalization of services and customer relationship management (CRM) [4]. When Speaker recognition is combined with speech recognition, it offers efficient interface between human natural language and computer for communication. In speaker identification and verification feature extraction approaches is most important one. Most common feature extraction techniques are Cepstral analysis [5, 6, 7 and 8] and Mel Frequency Cepstral Coefficients (MFCC) [9, 10 and 11]. Linear Prediction is a technique which is used as an intermediate method to derive the MFCC [5]. Perceptual Linear Prediction (PLP) is modification of LP technique which shows improved results but this technique is not widely used. Some other methods to derive the feature extraction approaches for speaker identification are Line Spectral Pairs (LSP) and Principal Spectral Components (PSC) [12].

This paper can be organized as follows: Literature Survey, Dual Tree-Complex Wavelet transform for feature extraction are described, in section II and section III and Speaker classification of Speaker identification by Relevance Vector Machine is described in section IV. Experimental results are given in section V.

Related Works

In speaker identification, feature extraction step converts the properties of the signal which are important for the pattern recognition task to a format that simplifies the distinction of the classes. The recognition step aims to estimate the general extension of the classes within feature space from a training set [24].

High-level features are generally related to a speaker's learned habits and style, such as particular word usage or idiolect. For humans, the information about the audio category is perceived by listening to a longer segment of audio signal. The information contained in the audio signal is the suprasegmental information. This information is the variation of the signal over long duration. In this case, speech is analyzed using the frame size and shift in the range of 50-200 ms. Studies made in [25, 26, 27 and 28] shows the significance of suprasegmental features in speaker recognition systems. These features are useful as their structure is not affected by the frequency characteristics of the transmission systems. Each of the four basic acoustic features of speech signal, i.e. pitch, intensity, duration and speech quality, is a carrier of a variety of types of linguistic, paralinguistic and non-linguistic information [29, 30].

Texts promoted speaker verification are verified by using algorithm of HMMs, though SVM, VQ and GMM are mostly used for text independent speaker recognition. In present days GMM is used as a classification for the speaker recognition system [13]. The GMM models the Probability Density Function (PDF) of a feature set as a weighted sum of multivariate Gaussian PDFs. It is equivalent to a single state continuous HMM, and may also be interpreted as a form of soft VQ [14].

Machine learning approach of SVM is used for speaker recognition and identification. These types of SVM classification are not insignificant compared to

approaches of GMM [15, 16]. But the combination of SVM and GMM provides the better improvement classification for the speaker recognition and verification [17].

In [18] the author compared two methods for speaker recognition using VQ classification with various HMM configuration. The continuous HMM provides better performance than discrete HMM. But, VQ technique works well on small training data set. Neural network has various structures used for speaker recognition and verification [19]. Various techniques for ANN such as Multi-Layer Perceptron (MLP) Networks, Radial Basis Function (RBF) Networks [20], Gamma Networks [21], and Time-Delay Neural Networks (TDNN) [22].

Different types of speech recognition system can be developed based on the type of speech, speaker and vocabularies used. These categories are used depending on the type of the application that the people use. Today's researches mainly focus on developing speech recognition systems for Indian languages [23].

Dual-Tree Complex Wavelet Transform (Dt-Cwt) For Feature Extraction

When compared to DT-CWT, Discrete Wavelet Transform (DWT) has some disadvantages such as oscillations, shift variance, aliasing and lack of directionality. Dual Tree Complex Wavelet Transform, a form of discrete wavelet transform which generates complex coefficients by using a dual tree of wavelet filters to obtain their real and imaginary parts. DT-CWT has the following properties to overcome the drawbacks of DWT:

- Approximate shift invariance;
- Good directional selectivity in 2-dimensions (2-D) with Gabor like filters also true for higher dimensionality (m-D)
- Perfect reconstruction
- Limited redundancy: 2× redundancy in 1-D (2d for d-dimensional signals), this is less than the log2N× redundancy of a perfectly shift-invariant DWT;
- Efficient order N computation. DT-CWT introduces limited redundancy (2m:1 form-dimensional signals) and allows the transform to provide approximate shift invariance and directionally selective filters by preserving the properties of perfect reconstruction and computational efficiency with balanced frequency responses. The main drawback of this transform is moderate redundancy.

The dual-tree complex DWT of a signal x(n) is implemented using two critically-sampled DWTs in parallel on the same data, as in Figure 1. To gain advantage over DWT, the filters designed in the upper and lower DWTs are different and are designed to interpret the sub band signals of the upper DWT as the real part of a complex wavelet transform, and lower DWT as the imaginary part. When designed in this way, the DT-CWT is nearly shifting invariant, in contrast to the classic DWT. The DT-CWT is used to implement 2D wavelet transforms where each wavelet is oriented, and useful for image processing such as image denoising and enhancement applications.

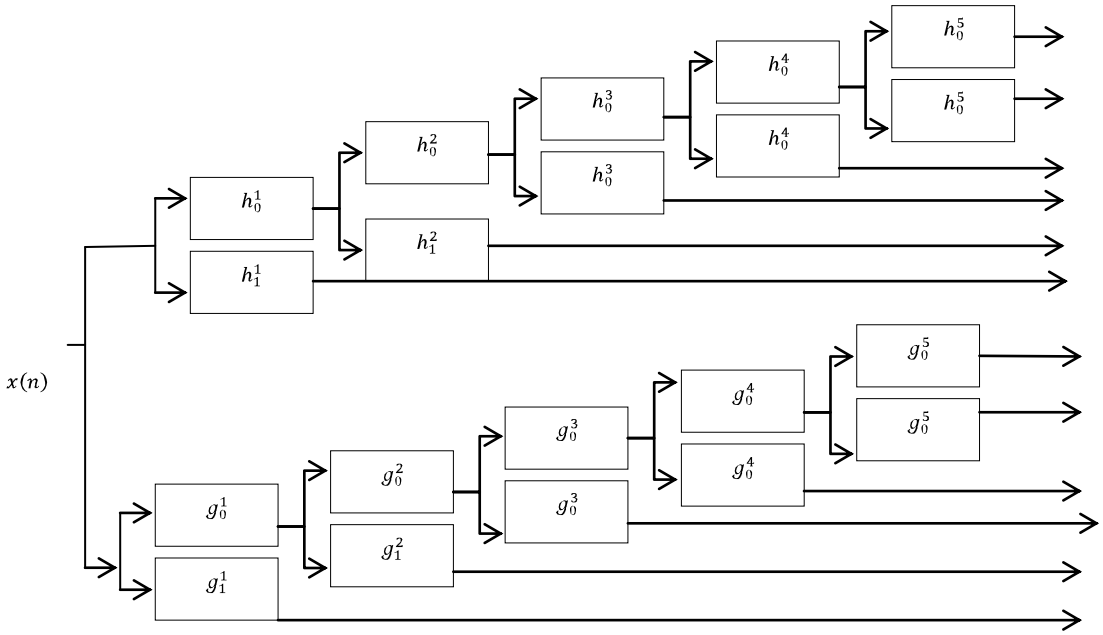

Figure 1: DT-CWT structure

There are two types of the 2D dual-tree wavelet transform, they are real and complex. The real 2-D dual-tree DWT is 2-times larger in space and the complex 2-D dual-tree DWT is 4-times expansive in space, and they are oriented in six distinct directions

Proposed Methodology

Speaker Verification Using Relevance Vector Machine

The Relevance Vector Machine (RVM) was used in [17] as a Bayesian counterpart to the SVM. Due to its simplicity and applicability, RVM made tremendous growth in the Machine Learning community. The RVM provides empirical Bayes treatment of function approximation by kernel basis expansion and attains a sparse representation of the approximating function by structuring a Gaussian prior distribution that implicitly creates a sparsity pressure on the coefficients appearing in the expansion. The use of independent Gamma hyperpriors produces product of independent marginal prior for the coefficients and hence it achieves the desired sparsity.

In order to minimize the dimensionality of the hyper parameter space, define a prior structure that affects the possibility of correlation between the hyper parameters of the coefficients distribution and hence it is possible to segregate a unique solution.

In this paper, RVM has been used for speaker identification. Relevance vector machine (RVM) is sparse linear model in which the basis functions are formed by a kernel function φ centred at the different training points:

$$y(x) = \sum_{i=1}^{N} W_i \varphi(x - x_i) \quad (1)$$

This model is similar as the support vector machines (SVM), the kernel function in the above equation does not meet the Mercer's condition and it needs φ to be a continuous symmetric kernel of a positive integral operator [18].

Multi-kernel RVM is an extension of the RVM model. It consists of different types of φ_m kernels and it is expressed as

$$y(x) = \sum_{i=1}^{m} \sum_{i=1}^{N} W_i \varphi_m (x - x_i) \quad (2)$$

The sparseness property enables choosing proper kernel automatically at each location by pruning all irrelevant kernels, hence it is possible that two different kernels remain on the same location.

Assume a two-class problem with training points $X = \{X_1, \ldots, X_N\}$ and corresponding class labels $t = \{t_1, \ldots, t_N\}$ with $t_i \epsilon \{0,1\}$. Applying the Bernoulli distribution [41], the likelihood (the target conditional distribution) can be expressed as:

$$p(t|W) = \prod_{i=1}^{N} \sigma\{(y(x_i))\}^{t_i} \left[1 - \sigma\{(y(x_i))\}^{1-t_i}\right] \quad (3)$$

Where $\sigma(y) -$ logistic sigmoid function

$$\sigma(y(x)) = \frac{1}{1+exp(-y(x))} \quad (4)$$

Consider α_i^* denotes the maximum a posteriori (MAP) estimate of the hyper parameterα_i. The MAP approximate for the weights is denoted by wMAP and it can be obtained by maximizing the posterior distribution of the class labels given the input vectors. It is equivalent to maximizing the objective of the function given by:

$$J(W_1, W_2, \ldots, W_N) = \sum_{i=1}^{N} logp(t_i|w_i) + \sum_{i=1}^{N} logp(w_i|\alpha_i^*) \quad (5)$$

Where the first term indicates the likelihood of the class labels and the second term indicates prior on the parameters W_i. Those samples associated with nonzero coefficients W_i which are called relevance vectors will contribute to the decision function.

The gradient of the actual function J with respect to w is given by:

$$\nabla J = -A^* W - \varphi^T (f - t) \quad (6)$$

Where$f = \left[\sigma(y(x_1)) \ldots . \sigma(y(x_1))\right]^T$, where φ have elements$\varphi_{ij} = K(x_i, x_j)$. The Hessian of J is

$$H = \nabla^2(J) = -(\varphi^T B \varphi + A^*) \quad (7)$$

Where $B = diag(\beta_1, \ldots, \beta_N)$ is a diagonal matrix with $\beta_i = \sigma(y(x_i))[1 - \sigma(y(x_i))]$

The posterior is approximated around W_{MAP} by a Gaussian approximation with covariance

$$\sum = -(H/w_{MAP})^{-1} \quad (8)$$

and mean is given by,

$$\mu = \sum \varphi^T Bt \tag{9}$$

RVM has several advantages which includes the number of relevance vectors can be much smaller than that of support vectors , RVM does not need the tuning of a regularization parameter (C) as in SVM during the training phase. Thus the proposed dataset can be classified using RVM classifier.

Experimental Result

PDA dataset [36] is used to evaluate proposed method. In the PDAs data set, the voice was recorded with a Compaq iPAQ 3630 built-in microphone and an Optimus Nova 80 close-talk microphone. The iPAQ data were recorded using the AD converter in the iPAQ, and the close-talk data were recorded using a Creative Sound board (except for speakers #1 & 2). Both channels were recorded with an 11.025 kHz sampling. Dataset contains 646 speech signals. Using this dataset, performance of Speaker verification can be analyzed using the false acceptance rate (FAR), the false rejection rate (FRR)

$$\text{FAR} = \frac{\#\text{ accepted imposter claims}}{\#\text{ imposter accesses}} \times 100\% \tag{10}$$

$$\text{FRR} = \frac{\#\text{ rejected genuine claims}}{\#\text{ genuine accesses}} \times 100\% \tag{11}$$

Comparison of FAR and FRR

Table 1: Comparison of Far and FRR

Techniques	**FAR (%)**	**FRR (%)**
MF-PLP	7.89	10.22
PLP	10.618	12.608
MFCC-SVM	7.24	10.25
MFCC-GMM	17.1	18.6
PLP-GMM	5.5	5.8
Proposed DT-CWT with RVM	4.32	4.12

The above table I provides the FAR and FRR for techniques of MF-PLP, MFCC-SVM, MFCC-GMM, PLP-GMM and Proposed method of DT-CWT with RVM. From the table, it is clearly observed that the proposed method of DT-CWT with RVM provides very low FAR and FRR of 4.32% and 4.12%. PLP-GMM [35] provides 5.5% FAR and 5.8% FRR, MFCC-GMM [35] provides 17.1% FAR and 18.6% FRR, MFCC-SVM [34] gives 7.24% FAR and 10.25% FRR and MF-PLP [33] provides 7.89% FAR and 10.22% FRR. From these result, we clearly understand that proposed technique gives low FAR and FRR values than other techniques.

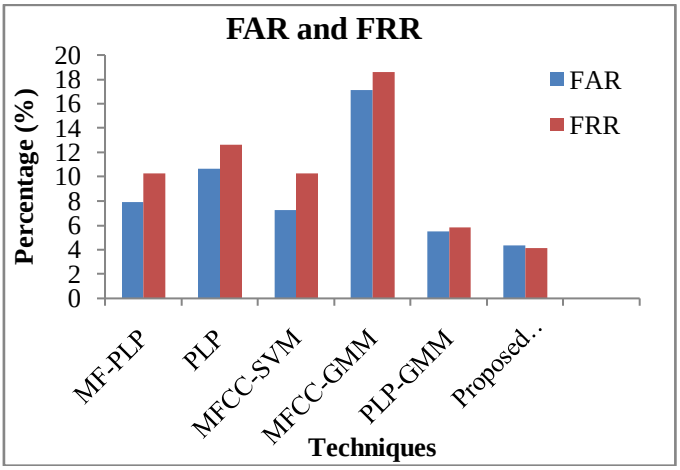

Figure 2: Comparison Between Far And Frr

The above Figure 2 shows the FAR and FRR for techniques of MF-PLP, MFCC-SVM, MFCC-GMM, PLP-GMM and Proposed method of DT-CWT with RVM. From the graph it clearly shows that the proposed method of DT-CWT with RVM provides the very less FAR and FRR compared to other approaches.

Execution Time

Table 2: Execution Time

Techniques	**Execution time (sec)**
MFCC	12
Hybrid Method of AM demodulation and wavelet filters	11
Proposed DT-CWT with RVM	8

The above table II provides the comparison of execution time for MFCC [31], Hybrid method [32] and proposed DT-CWT with RVM. From the table, it is observed that the proposed method gives less execution time of 8 seconds where Hybrid AM demodulation and wavelet filter gives 11 seconds and MFCC takes 12 seconds.

Speaker Identification Rate

Table 3: Speaker Identification Rate

Techniques	**Speaker identification rate(%)**
MFCC with GMM	77.36
IMFCC (Inverted MFCC)	77
Kullback-Leibler divergence	93
Proposed DT-CWT with RVM	95

The above table III provides the comparison of speaker recognition rate between existing approaches such as MFCC and GMM [35], IMFCC [38], Kullback-Leibler divergence [40] and proposed DT-CWT with RVM. From the table, the speaker identification rate of MFCC with GMM provides accuracy of 77.36%, IMFCC (Inverted MFCC) [38] for polycost database using triangular filter gives 77%, Kullback-Leibler divergence [40] for different gender provide the accuracy of 93%. But the proposed DT-CWT with RVM gives 95% of identification rate which is better than existing techniques.

Conclusion

This paper provides the speaker verification using RVM. In this paper, speech signals are extracted by using DT-CWT. Compared to other wavelet transform, this method of DT-CWT provides the better feature extraction. The standard CWT has the disadvantages such as shift sensitivity and poor directionality. The DT-CWT overcomes the standard CWT issues and has advantage of wide range of directionality. The proposed method DT-CWT with RVM provides better accuracy, less execution time when compared with other approaches.

References

[1] R. Summerfield, T. Dunstone, C. Summerfield, “Speaker Verification in a Multi-Vendor Environment”,www.w3.org/2008/08/siv/Papers/Centrelink/w3c-sv_multivendor.pd

[2] T. Becker, M. Jessen, C. Grigoras, “Forensic Speaker Verification Using Formant Features and Gaussian Mixture Models”, Interspeech 2008.

[3] M. Sokolov, “Speaker verification in the World Wide Web”, Eurospeech 1997.

[4] J. Markowitz, “Using speech recognition in customer relationship management to be more effective”, In DCI customer relationship management conference, 1999.

[5] S.I Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Transactions on Acoustics Speech and Signal Processing, vol. 29(2), pp. 254-72, Apr. 1981.

[6] K.T. Assaleh, R.J. Mammone, “Robust cepstral feature for speaker identification.”, In proceedings of the IEEE ICASSP, Vol 1, pp. 129-132, April 1994.

[7] K.T. Assaleh, “Supplementary orthogonal cepstral features”, In ICASSP 1995.

[8] R. Sethuraman, J.N.Gowdy, “A cepstral based speaker recognition system”, In 21st Southeastern symposium on System Theory, 1989.

[9] D. Ververidis, C. Kotropoulos, "Gaussian mixture modeling by exploiting the mahalanobis distance", IEEE transactions on signal processing, Vol. 56, No. 7, pp. 2797-2811 , July 2008.

[10] K.N. Stevens, "Sources of inter and intra-speaker variability in acoustic properties of speech sounds", In 7th international congress on phonetic sciences, 1971, Montreal, Canada.

[11] B. Yegnanarayana, S.R.M Prasanna, J.M. Zachariah, C.S. Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system", IEEE Transactions on Speech and Audio Processing, Vol. 13, No. 4, pp. 575-582, July 2005

[12] M.M. Homayounpour, G. Chollet, "A comparison of some relevant parametric representation for speaker verification", In ESCA Workshop on automatic speaker recognition, Identification and verification.

[13] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", Speech Communication, Vol.17, pp.91-108, 1995.

[14] X. Zhu, et al., "A comparative study of mixture-Gaussian VQ, ergodic HMMs and left-to-right HMMs for speaker recognition", In international symposium on Speech, Image processing and Neural Networks, 1994.

[15] W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, and P.A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," Computer Speech and Language , vol. 20, pp. 210-29, 2006.

[16] V. Wan, and S. Renals, "Speaker verification using sequence discriminant support vector machines," IEEE Transactions on Speech Audio Processing, vol. 13, pp. 203-10, 2005.

[17] S. Bengio, J. Mariethoz, "Learning the decision function for speaker verification", In IEEE International conference on Acoustics, Speech and Signal Processing, 2001.

[18] T. Matsui, and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and Discrete/continuous HMMs," IEEE Transactions on Speech Audio Processing, vol. 2(3), pp. 456-9, July 1994.

[19] E. Bozkurt, E. Erzin, C.E. Erdem, A.T. Erdem, "Automatic emotion recognition for facial expression animation from speech", 17th IEEE conference on signal processing and communications, pp. 989-992, 2009.

[20] C.M. Bishop, Neural Networks for pattern recognition, 1995, Oxford University Press.

[21] C. Wang, D. Xu, C.P. Jose, "Speaker verification and identification using gamma neural networks", In international conference on neural networks, 1997.

[22] X. Wang, "Text dependent speaker verification using recurrent neural time delay neural networks for feature extraction", In IEEE signal processing workshop, 1993.

[23] M. Chandrasekar, M. Ponnavaikko, "Spoken TAMIL Character Recognition", in Electronic Journal Technical Acoustics (EJTA), ISSN 1819-2408, 2007.

[24] A.O. Afolabi, A. Williams, and O. Dotun, "Development of a text dependent speaker identification security system", Research Journal of Applied Sciences, 2 (6), pp. 677-684, 2007.

[25] B. Yegnanarayana, S.R.M. Prasanna, J.M. Zachariah, and C.S. Gupta,, "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system," IEEE Trans. Speech Audio Process. , vol. 13(4), pp. 575-82, July 2005.

[26] Grazyna Demenko, "Analysis of suprasegmental features for speaker verification", 8th Australian International Conference on Speech Science and Technology, 2000

[27] B. Yegnanarayana, K. Sharat Reddy, and S.P. Kishore, "Source and system features for speaker recognition using AANN models," in proc. Int. Conf. Acoust., Speech, Signal Process. , Utah, USA, Apr. 2001.

[28] P.D. Bricker, "Statistical techniques for talker identification", Bell Svst. Techn. Jour. Vol.50 April 1971

[29] B. S. Atal "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification "J. Acoust. Soc. Am. Volume 55, Issue 6, pp. 1304-1312 (1974)

[30] Tommy Kinnuen, "Spectral Features for Automatic Text-Independent Speaker Recognition " thesis University of Joensuu ,Dec.2003.

[31] P. Mermelstein and S. Davis, "Comparison of parametric representation for mono syllabic word recognition in continuously spoken sentences", In IEEE Transactions on Acoustic Speech and Signal Processing, Vol. 28, No. 4, pp. 357-366, 1980.

[32] [32]Vibha Tiwari and Jyoti Singhai, "Wavelet Based Noise Robust Features for Speaker Recognition", Gyan Ganga Institute of Technology and management Bhopal, India.

[33] A.Revathi, R.Ganapathy and Y.Venkataramani, "Text Independent Speaker Recognition and Speaker Independent Speech Recognition Using Iterative Clustering Approach", International Journal of Computer science & Information Technology (IJCSIT), Vol 1, No 2, November 2009.

[34] Shi-Huang Chen and Yu-Ren Luo, "Speaker Verification Using MFCC and Support Vector Machine", Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong.

[35] R. Rajalakshmi and A. Revathy, "Comparison of MFCC and PLP in Speaker Identification using GMM", International Conference on Computing and Control Engineering (ICCCE 2012), 12 & 13 April, 2012.

[36] http://www.speech.cs.cmu.edu/databases/pda/

[37] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus

for large vocabulary continuous speech recognition research," J. Acoust. Soc. Jpn. (E), vol. 20, no. 3, pp. 199–206, 1999.

[38] S. Chakroborty and G. Saha, "Improved Text-Independ-ent Speaker Identification Using Fused MFCC and IMFCC feature Sets Based on Gaussian Filter," Interna-tional Journal of Signal Processing, Vol. 5, No. 1, 2009, pp. 11-19.

[39] Seiichi Nakagawa, Longbiao Wang, and Shinji Ohtsuka, "Speaker Identification and Verification by Combining MFCC and Phase Information", IEEE Transactions on Audio, speech, and language processing, vol. 20, no. 4, may 2012.

[40] R. Saeidi, P. Mowlaee, T. Kinnunen and Z. H. Tan, "Sig-nal-to-Signal Ratio Independent Speaker Identification for Co-Channel Speech Signals," Proceedings of Interna-tional Conference on Pattern Recognition (ICPR2009), 2009, pp. 4565-4568

[41] Weisstein, Eric W. "Bernoulli Distribution." From MathWorld--A Wolfram Web Resource.