ISSN 2278-3091

Volume 9, No.1, January – February 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse77912020.pdf

https://doi.org/10.30534/ijatcse/2020/77912020

Speaker Feature Modeling Utilizing Constrained Maximum Likelihood Linear Regression and Gaussian Mixture Models

Elmer R. Magsino

ECE Department, De La Salle University, Manila, Philippines, elmer.magsino@dlsu.edu.ph

ABSTRACT

This paper describes a speaker recognition system based on feature extraction utilizing the constrained maximum likelihood linear regression (CMLLR) speaker adaptation, while using Gaussian mixture models (GMM) to model the speaker and background models. For the input acoustic signals, the cepstral features are derived to highlight the differences between test and training utterances. The CLSU dataset is used to test the efficiency and performance of the proposed CMLLR, Support Vector Machine, and GMM methods for modeling the speaker's voice by characterizing the speaker features.

Key words: Constrained Maximum Likelihood Linear Regression, Speaker Recognition

1. INTRODUCTION

Speaker recognition is a field of study that concerns in knowing people from their voices and accurately answers the question: "Who is speaking?" Recognizing a speaker relies on the unique acoustic features of each individual. These acoustic patterns characterize the throat's and mouth's shape and size, as well as the speaker's behavioral patterns, e.g., pitch, speaking style, etc. [1].

Two applications of speaker recognition can be distinguished, namely: (1) Verification or Detection and (2) Identification. Speaker verification solves the problem of verifying a person's identity using voice. Speaker identification deals in identifying a person solely by their voice. The difference is that: the identity of a speaker is already known from the former application and needs to be verified while the latter deals with determining an unknown speaker's identity, thus identity has not yet been established [2]. In recognizing a speaker's voice, two types of constraints that are used, namely, text-dependent [3] and text-independent [4]. The former relies on a given set of words to be uttered by a speaker, while the latter allows any words to be spoken which is advantageous to avoid fraud [5]. Other works have considered the physical constraints such as trunk movement and physical constraint in the production of speech [6], multiple noisy environments [7], and even language translations [8]. Convolutional neural network can also be used in speaker recognition just as it is employed in face recognition system [9]. Related to this field is the blind source separation or the cocktail party effect wherein a specific speaker's voice can be extracted from a group of noisy or unwanted environment [10]. Neural networks can also be trained to model speaker features as what was done in [11] and other applications [11].

In speaker modeling, background speaker and universal background modeling methods can be utilized. In this work, we adapt the constrained maximum likelihood linear regression (CMLLR) [11] and support vector machines (SVM) methods for recognizing a person's voice. Gaussian mixture models (GMM) models the speaker and its background. The major contributions of this work are summarized below. Background modeling employs several speakers to cover the space of the alternative hypothesis [11], while the universal background modeling has a pool of speakers in a single utterance and then train a single model. This is the predominant approach in modern speaker recognition systems [12]. Nowadays, i-vectors methodology is used because they perform very well when both the training and test datasets are highly the same [13]. An extended application of speaker recognition involves multi-party conversation. Aside from recognition, speaker classification can predict the utterance of a speaker [14].

This paper is organized as follows. Section 2 discusses theories related to CMLLR. Section 3 outlines the feature extraction, statistical modeling and speaker adaptive training using CMLLR phases. Experimental results and discussions are provided in section 4. Section 5 concludes the work and gives some recommendations.

2. CMLLR FOR SPEAKER RECOGNITION

In this section, we present the CMLLR method on how to recognize a speaker. Maximum-likelihood linear regression (MLLR) [15] finds the optimal affine transformation of a model. In speaker adaptation, employing linear transformations is common due to the small amount of needed data for modeling systems even if the system is large. However, linear transformations are limited because it is model based. In an environment with many possible sources of voice, it is advantageous if an adapted model can be used for multiple speakers, instead of developing new models for new speakers [16].

CMLLR is a model-space transformation but uses the same transformation matrix to adapt the model means and covariances. With this assumption, adaptation can be done in the feature space rather than model space. The affine transformation used in CMLLR is given by:

$$\hat{\mu} = A\mu + b$$
$$\hat{\Sigma} = A\Sigma A^T \quad (1)$$

The transformation matrix A is also estimated using the Expectation-Maximization step.

Features are transformed as:

$$\hat{o}(\tau) = Ao(\tau) + b = W\zeta(\tau) \quad (2)$$

where A is the transformation matrix and b the constant bias. $W = \begin{bmatrix} b^T & A^T \end{bmatrix}^T$ is the extended transformation matrix and $\zeta(\tau) = \begin{bmatrix} 1 & o(\tau) \end{bmatrix}$ is the extended observation vector at time τ.

3. METHODOLOGY

This section discusses the general architecture of the speaker recognition system, cepstral feature extraction, statistical modeling and speaker adaptive training.

3.1 General Architecture

Figure 1 shows the general architecture of the top module of the speaker verification system in which it is divided into two stages: (1) offline training and (2) testing.

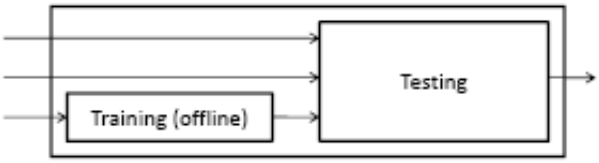

Figure 1: Speaker Verification System General Architecture

The offline training implementation block diagram is shown in Figure 2. A given speech signals will serve as training data wherein speech parameters, features and observations will be extracted. From these derived data, statistical modeling based on Gaussian Mixture Modeling will be done.

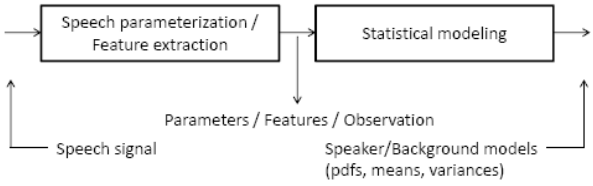

Figure 2: Offline Training Block Diagram

3.2 Cepstral Feature Extractions

The data used are from CLSU Speaker Recognition Corpus [15]. The objective is to determine whether a certain segment is uttered by the target speaker or not. Cepstral feature vectors consist of 15 MEL-PLP cepstrum coefficients, 15 Δ coefficients and its energy, 15 ΔΔ coefficients and its energy. Cepstral feature extraction highlights the acoustical differences between segmented parts of the test utterance [16].

3.3 Statistical Modeling

Given that the input speech signal is a random process, the speaker recognition system must be able to make decisions based on statistical models of the expected input. Gaussian Mixture Models are used for modelling the input signals because the linear combination of Gaussian basis functions can represent a wide range of sample distributions

3.4 Speaker Adaptive Training

The testing phase is done as shown in Figure 3. Speaker adaptive training (SAT) uses CMLLR alongside with Support Vector Machines (SVM) [17] which is a form of discriminative-based leaning technique. CMLR provides tighter coupling between background and speaker models. SVM, on the other hand, takes the high-dimensional feature per speaker and separates the impostors and the clients of the system. The advantage to this technique is that scoring lies using only the inner product between the target model (new feature technique) and the GMM supervector.

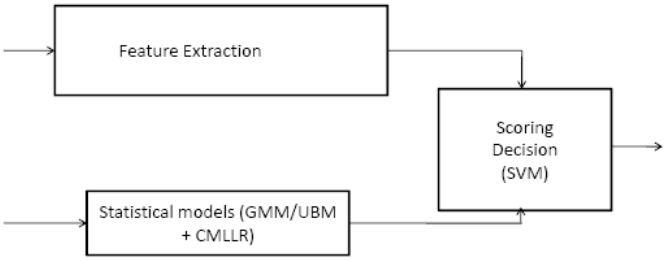

Figure 3: Testing Phase Block Diagram

4. RESULTS AND DISCUSSION

We present below the training and testing results of the CMLLR-SVM methods below.

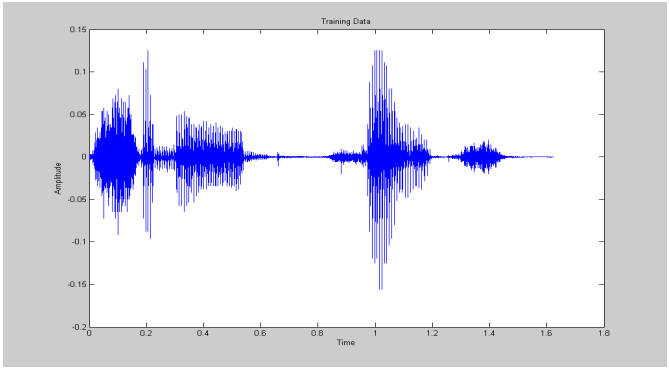

Figure 4: Sample speech test data

4.1 Cepstral Feature Extraction

A sample of test data being used to train the system is shown in Figure 4. The speech signal has a length of 1.6225 seconds; its sampling frequency is 8000 Hz. The dimension derived from the speech signal is 161 frames. This is assumed that a 30ms Hamming window is applied to every 10ms of the speech signal and neglecting the first sample. Thus, the signal has 47 – 161 features.

Shown in Figure 5 are the 47 features extracted from the above signal. The last two peaks of the plot show the Δ and ΔΔ energy of the test signal respectively.

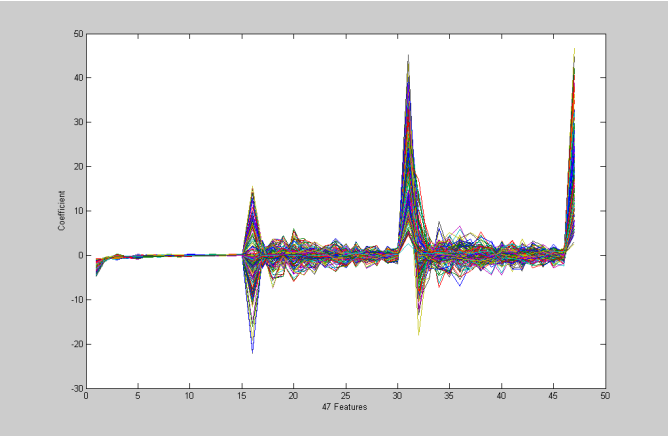

Figure 5: 47 Features extracted from sample speech test data

Figure 6 exhibits the per feature characteristic of each of the 161 frames. The red and violet plots are the Δ and ΔΔ energy plots, respectively. Majority of the per feature of the speech signal is close to zero.

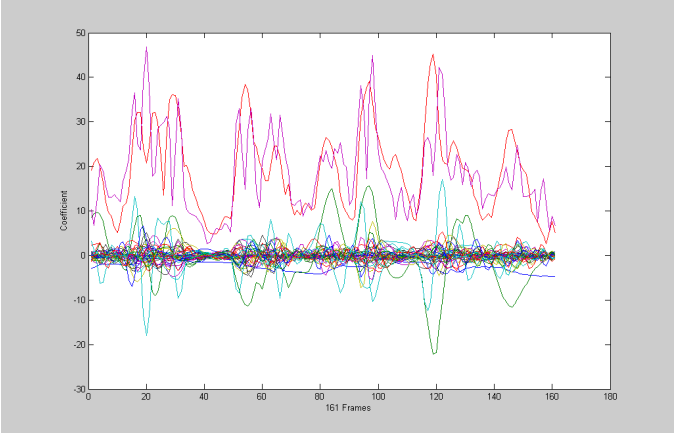

Figure 6: Per Feature plot from the sample speech test data

4.2 Statistical Modeling

Once the features have been extracted from a given test speech signal, these are now used to derive a Gaussian Mixture Model (GMM) wherein the mean and variance of each Gaussian are derived [12]. To complete the modeling, the weight of each of Gaussian that comprises the mixture model is also derived. The Expectation-Maximization algorithm (EM) [19] is used to determine these parameters of weights, means, and variances. In modeling the GMM, the usual mixture order used for speaker modeling varies from 64 – 256 while 512 – 2048 is usual for background modeling. In this experiment, the mixture order used is 256. Seven iterations are used to arrive at the estimated weights, means and variances of the GMM.

Shown in Figure 7 are the estimated weights of each of the 256 Gaussian models. The sum of all the weights is equal to one.

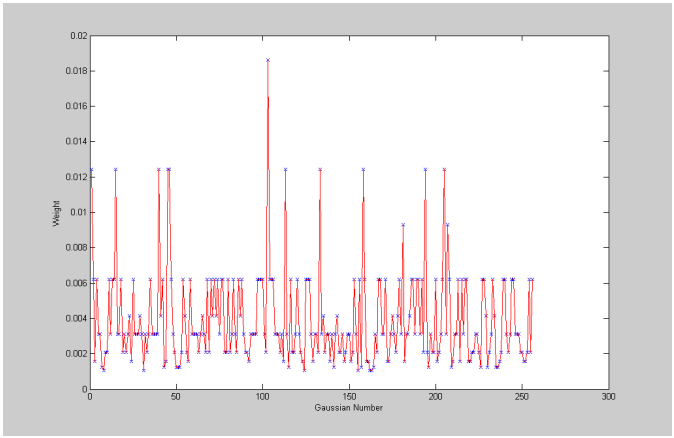

Figure 7: Estimated weights of each of the 256 used Gaussian Models

Figure 8 shows the mean modeling for the 256 GMM for each feature. The peaks show the estimated mean for the Δ and ΔΔ energy.

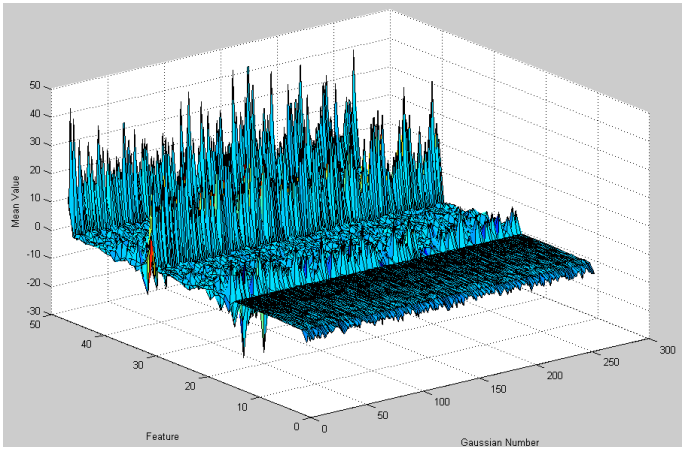

Figure 8: Estimated mean of each of the 256 used Gaussian Models

Figure 9 shows the variance modeling for the 256 GMM for each feature. During estimation, once a computed variance is less than 0.01, it is automatically clipped to 0.01.

The total training latency requires approximately five hours of speech training for an hour of input signal.

5. CONCLUSION AND FUTURE WORK

This work has successfully shown the successful feature extraction, statistical modeling, and speaker adaptation based on constrained maximum likelihood regression (CMLLR). The Gaussian Mixture Modeling (GMM) is also shown to effectively estimate the feature distribution of each speech frame. Finally, expectation-maximization algorithm is implemented to estimate the necessary process parameters. In the future, SVM training and testing will be evaluated to finally complete the speaker recognition system.

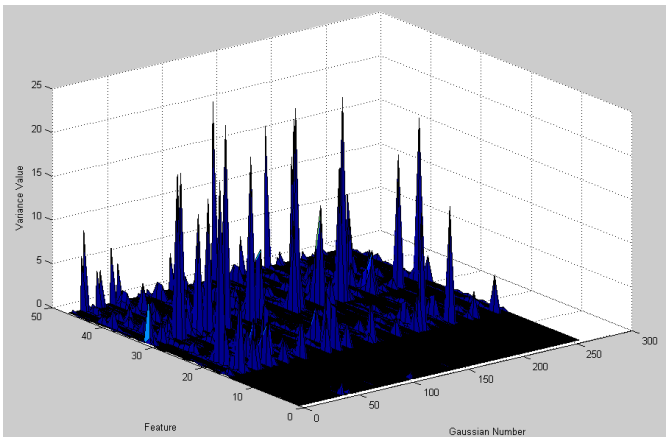

Figure 9: Estimated variance of each of the 256 used Gaussian Models

REFERENCES

[1] P. Ehkan, F. F. Zakaria, M. Warip, Z. Sauli and M. Elshaikh, "Hardware implementation of MFCC-based feature extraction for speaker recognition," in *Advanced Computer and Communication Engineering Technology*, Springer, 2015, pp. 471-480. https://doi.org/10.1007/978-3-319-07674-4_46

[2] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74-99, 2015.

[3] F. Chowdhury, Q. Wang, I. L. Moreno and L. Wan, "Attention-based models for text-dependent speaekr verification," arXiv preprint arXiv:1710.10470, 2017.

[4] G. Bhattacharya, J. Alam and P. Kenny, "Deep speaker embeddings for short-duration speaker verification," in *Inter-speech*, 2017.

[5] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12-40, 2010. https://doi.org/10.1016/j.specom.2009.08.009

[6] K. Sakai, T. Minato, C. Ishi and H. Ishiguro, "Speech Driven Trunk Motion Generating System based on Physical Constraint," in *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, New York, 2016.

[7] Z. Wang, E. Vincent , R. Serizel and Y. Yan, "Rank-1 constrained multichannel Wiener filter for speech recognition in noisy environments," *Computer Speech & Language*, vol. 49, pp. 37-51, 2018.

[8] M. M. Bailon, M. C. De Silva, R. P. Lapuz, J. L. A. Tinio, T. B. K. Yu and R. C. Gustilo, "Filipino to chinese speech-to-speech translator using neural network with database system," *International Journal of Emerging Trends in Engineering Research*, vol. 7, no. 9, pp. 276-282, 2019.

[9] J. Del Rosario, "Development of a face recognition system using deep convolutional neural network in a multi-view vision environment," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 3, pp. 369-374, 2019. https://doi.org/10.30534/ijatcse/2019/06832019

[10] C. Sandiko and E. R. Magsino, "A Blind Source Separation of instantaneous acoustic mixtures using Natural Gradient Method," in *2012 IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2012*, Penang, Malaysia, 2012.

[11] M. S. Rozario, A. Thomas and D. Mathew, "Performance Comparison of Multiple Speech Features for Speaker Recognition using Artifical Neural Network," in *2019 9th International Conference on Advances in Computing and Communication (ICACC)*, 2019.

[12] A. D. Africa, L. R. Bulda, M. Z. Marasigan and I. F. Navarro, "A Study on Number Gesture Recognition using Neural Network," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 4, pp. 1076-1082, 2019. https://doi.org/10.30534/ijatcse/2019/14842019

[13] S. Sivaram, K. Santhosh and A. Kumar, "Enhancement of dysarthric speech for developing an effective speech therapy tool," in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, 2017.

[14] A. K. Sarkar and Z.-H. Tan, "Incorporating pass-phrase dependent background models for text-dependent spekar verification," *Computer Speech & Language*, vol. 47, pp. 259-271, 2018.

[15] J. Chang and D. Wang, "Robust speaker recognition based on DNN/i-Vectors and speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[16] Q. Wang, W. Rao, S. Sun, l. Xie, E. Chng and H. Li, "UNSUPERVISED DOMAIN ADAPTATION VIA DOMAIN ADVERSARIAL TRAINING FOR SPEAKER RECOGITION," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[17] Z. Meng, L. Mou and Z. Jin, "Towards Neural Speaker Modeling in Multi-Party Conversation: The Task, Dataset, and Models," in *The Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[18] P. Dunn and G. Smyth, "Beyond Linear Regression: The Method of Maximum Likelihood," in *Generalized Linear Models With Examples*, New York, Springer, 2018.

[19] A. Martin and M. Przybocki, "Speaker recognition in a multi-speaker environment," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[20] T. Lander, "CLSU: Foreign accented english release 1.2," in *Linguistic Data Consortium*, Philadelphia, 2007.

[21] I. Trabelsi and D. Ayed, "On the use of different feature extraction methods for linear and non linear kernels," in *2012 6th international conference on sciences of electronics, technologies of information and telecommunications (SETIT)*, 2012.

[22] T. Yu and K. T. Nwet, "Myanmar News Sentiment Analyzer using Support Vector Machine Algorithm," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 6, pp. 3520-3525, 2019.

[23] D. A. Reynolds and R. C. Rose, "Robust Text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72-8, 1995.

[24] T. Moon, "The expectation-maximization algorithm," *IEEE Signal processing magazine* , vol. 13, no. 6, pp. 47-60, 1996.
https://doi.org/10.1109/79.543975