International Journal of Engineering Research
Volume No.5, Issue No.5, pp : 347-352

ISSN:2319-6890)(online),2347-5013(print)
1 May 2016

On the Use of Gaussian Mixture Model (GMM) Technique and YOHO Corpora for Automatic Speaker Recognition for Nigerian Tribal Languages

[1]Afolabi Lateef Olashile, [2*]Ehiagwina Ojiemhende Frederick, [3]Onaowola Hassan Jimoh, [4]Abubakar Nafiu Sidiq,[5]Seluwa Oludare Emmanuel
[1,2,5] Department of Electrical Engineering, [3,4] Department of Computer Technology Engineering
Federal Polytechnic, Offa, Offa Kwara State, Nigeria
[1]mrshile@yahoo.com, [2]*Frederick.ehiagiwna@fedpoffaonline.edu.ng, [3]honawola@yahoo.com
[4]assidiq123@yahoo.com, [5]seluwaoe2014@gmail.com
[*]*corresponding Author*

Abstract: ***Many levels of information can be gotten from speech such as the message being spoken, the language been spoken, information about the speaker and the emotional state of the speaker. Several literature have reported on automatic speaker recognition system. This article overviewed speaker recognition systems with emphasis on those using Gaussian Mixture Model (GMM). Subsequently, via a statistical based speaker-modeling technique that represents the underlying characteristic sounds of a person's voice an analysis of the speech of speakers from Hausa, Yoruba and Igbo tribes in Nigeria was performed. Speaker recognizers that are capable of recognizing a speaker that is text-independent was designed. The wavelet toolbox of MATLAB® 2007 was used. Performance of the systems is evaluated for a wide range of speech quality; from clean speech to cell-phone speech, by using YOHO standard speech corpora. An Error Rate (ERR) of false acceptance rate of 0.51% and a false rejection rate of 0.65% at a 0.1% false-acceptance rate were obtained. ERR of compensation channel of 0.96%. Same experiment in identification section was also conducted, consequently 0.28% identification error rate was achieved. Using 4.45% of the speaker utterance 0.7% identification error rate was achieve.***

Keyword— Decision Threshold, Gaussian mixture model (GMM), Mel-scale filter, Speaker recognition, YOHO-corpus

I. INTRODUCTION

Automatic speaker recognition (ASR) is an aspect of biometric engineering. Biometric involves using metric related to human traits/features for the purpose of identification, authentication and access control. Traits uniquely identifying individuals could be physiological or behavioural. The behavioural characteristics among other traits include voice [1, 2, 3, 4, 5]. Automatic speaker recognition attempts to design algorithms that exploit the discriminative features of the human speech as a behavioural traits. It also exploits the uniqueness of the anatomy of the vocal tract for access control [6]. The algorithm is implemented on a computer so as to detect, identify and recognize the voice of a speaker. There are so many factors contributing to the growth and development of biometric speaker recognition systems, such as; the discovering of biometric traits that more accurately convey speaker dependent information. This among other factors has led researchers to progressively develop sophisticated speaker recognition algorithm. Over the past 70 years there have been improvements in the performance of these algorithm in more realistic evaluation speech corpora [7, 8, 9]. The aim of ASR is to extract, characterize and subsequently recognize the information about the speaker [10, 6].

As of 2015, specialized automatic speaker recognition system have been deployed in the following industries: telecommunications, banking transactions, crime agencies, and so on [11, 3]. Verification forms the basis for most speaker-recognition applications. As earlier present in [12], trendy applications include: computer log-in, telephone banking. Other applications are in calling cards, and cellular-telephone fraud prevention substitute or supplement a memorized personal identification code with speaker verification. Verification can also be applied as an information retrieval tool for retrieving messages from voice mailboxes.

Speaker-recognition tasks are further distinguished by the constraints placed on the text of the speech used in the system. In a *text-dependent system*, the spoken text used to train and test the system is constrained to be the same word or phrase. However, in contract, in a *text-independent system*, training and testing speech is completely unconstrained. This system provides more flexibility than the text-dependent system because pass phrases used by claimants can be changed regularly without retraining to help thwart an impostor with a tape recorder. Therefore, ASR involves training and testing phases. The process of familiarizing the biometric system with the voice characteristics of the speaker registering is referred to as training phase. While the testing phase is the actual recognition [6].

The aims of the paper is to use Gaussian mixture speaker model (GMM) to optimize the performance of automatic speaker recognition by reducing the error rate (ERR) between the real claimant and imposter in terms of speakers of *Hausa*, *Yoruba* and *Igbo* languages in Nigeria. Speaker verification requires distinguishing a speaker's voice known to the system from a potentially large group of voices unknown to the system.

Speakers known to the system who claim their true identity are called *claimants;* speakers, either known or unknown to the system, who pose as other speakers are called *impostors.* There are two types of verification errors:

a. false acceptances-the system accepts an impostor as a claimant (false-acceptance errors)

b. false rejections-the system rejects a claimant as an impostor (false-rejection error)

Many challenging problems and limitations remain to be overcome in recognizing speaker or to detect the true claimer from imposter.

i. In speaker identification, the unknown voice is assumed to be from the predefined set of known speakers.

ii. The difficulty of identification generally increases as the speaker set (or speaker population) increases.

Applications of pure identification are generally unlikely in real situations because they involve only speakers known to the system, called enrolled speakers. Figure 1 shows the schematic diagram of applications of speaker recognition.

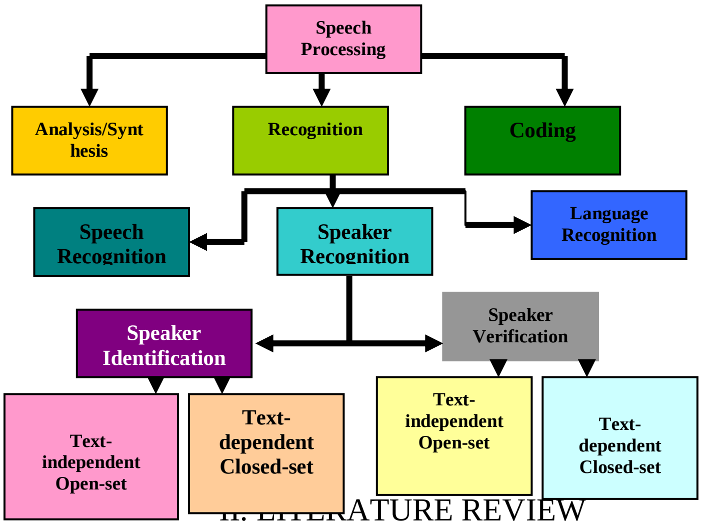

Fig 1: Schematic diagram of applications of speaker

II. LITERATURE REVIEW

ed

ne

nonverbal, when recognizing speakers. These cues are not well understood, but range from high-level cues, which are related to semantic or linguistic aspects of speech, to low-level cues, which are related to acoustic aspects of speech. Speaker verification has co-evolved with the technologies of speech recognition and speech synthesis because of the similar characteristics and challenges associated with each. And as presented in [13, 14, 15, 16], Gunnar Fant, a Swedish professor, published a model describing the physiological components of acoustic speech production, based on the analysis of x-rays of individuals making specified phonic sounds. Later, in 1970, Dr. Joseph Perkell used motion x-rays and included the tongue and jaw to expand upon the Fant's model [16]. Whereas, [17, 18] examined the following: mel-frequency and linear-frequency filter bank cepstral coefficients and perceptual linear prediction (PLP) cepstral coefficients.

Earlier speaker recognition systems used the average output of several analog filters to perform matching, often with the aid of humans-in-the-loop (HILT). In 1976, Texas Instruments built a prototype system that was tested by the U.S. Air Force and The MITRE Corporation. In the mid-1980s, the National Institute of Standards and Technology (NIST) developed the NIST Speech Group to study and promote the use of speech processing techniques. Moreover, an experimental evaluation of different features and channel compensation methods for robust speaker recognition was presented in [19]. In that presentation maximum likelihood classifier based on Gaussian mixture densities. Additionally, in that same year how to improve voice identification using nearest-neighbor distance measure (NNDM) was presented in [20, 21].

Since 1996, under funding from the National Security Agency, the NIST Speech Group has hosted yearly evaluations, the NIST Speaker Recognition Evaluation Workshop, to foster the continued advancement of the speaker recognition community [22, 23, 21]. In this paper Gaussian Mixture Model (GMM) will be reviewed and applied. Hence, the following presents a review of some application of GMM. Some useful discourses respecting speaker recognition are found in [10, 11, 24, 25, 26].

An extension of GMM using independent component analysis that can improve classification in comparison with standard GMM was presented in [27]. It is applied where source components have non-Gaussian densities. Moreover, [28] presented a method which improves adaptive background mixture model. It reinvestigated the update equations, and utilize different equations at different phases. This allows the developed system learn faster and more accurately as well as adapt effectively to changing environments. When incorporate with the shadow detection, the method results in an improved segmentation. On the other hand [29] presented a GMM as a bases of tokenization in language identification. About six years later, a proposal for an efficient approach to search for the global threshold of image using GMM was given by [30]. A demonstration of a statistical representation of distribution system loads using GMM on a 95-bus generic distribution network model was shown in [31]. However, improvement of the method bases on a priori knowledge is required. Furthermore, a presentation in which an all Hidden Markov Model (HMM) states share structure similar to GMM with the same number of Gaussians in each state was given in [32]. It observed that the developed model (equation 1) is compatible with most standard models. The distribution that produces a vector within HMM state pair m, j is a GMM:

$$p\left(x\middle|j\right)=\sum_{m=1}^{M_j} C_{jm} \sum_{i=1}^{m_j} w_{mji} N\left(x:\mu_{mji},\sum{}_i\right) \qquad (1)$$

where j index represents individual context-dependent phonetic states, with $1 \le j < J$, for covariance matrices $\sum_{ji}$ which are shared between states, mixture weight w_{mji} and mean μ_{mji} . x is D-dimensional feature vector, M stands for the number of uni-modal Gaussian densities [33]

$$w_{jmi} = \frac{\exp w_{jmi}{}^{T} V_{jmi}}{\sum_{i=1}^{I} \exp w_{jmi}{}^{T} V_{jmi}} \qquad (2)$$

Speech corpora is used primarily because it permits speech researchers to compare performance of different techniques using same data. Therefore, it is easier for researchers to know approaches needing further studies. Also, standard corpora can be used to measure current state-of-the-art performance in research areas for particular tasks and note shortcomings that need further research. Much description of various speech corpora is found in [34]. Other descriptions of one or more speech corpora are found in [35, 36, 37, 38, 39] This research uses the YOHO corpus, which is widely reported in the literature [40, 34, 41, 42] is designed to support text-dependent speaker verification evaluation for Government secure access applications, because it suits the environment of where the speech are collected.

III. METHODOLOGY

Most automatic speaker-recognition systems rely upon spectral differences to discriminate speakers. Natural speech is not simply a concatenation of sounds. Instead, it is a blending of different sounds, often with no distinct boundaries between transitions.

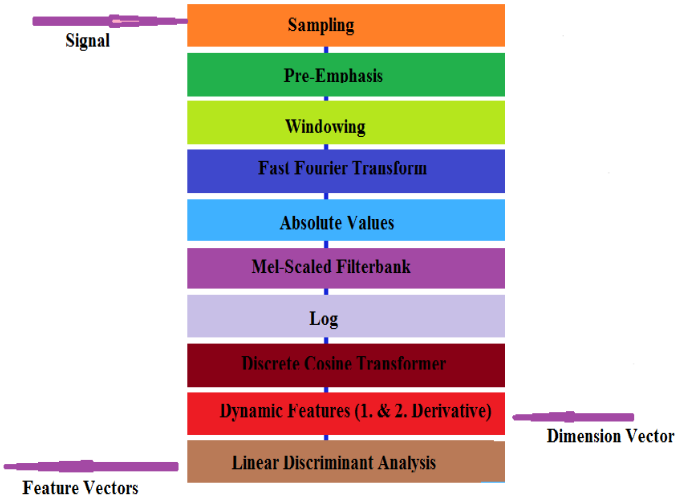

Fig 2: Schematic diagram of vector feature extraction

To obtain steady-state measurements of the spectra from continuous speech, we perform short-time spectral analysis, which involves several processing steps, as shown in Figure 2

1. The speech is segmented into frames by a 20-msec window progressing at a 10msec frame rate.
2. A speech activity detector is then used to discard silence and noise frames.
3. For text-independent speaker recognition, removing silence and noise frames from the training and testing signals is important in order to avoid modeling and detecting the environment rather than the speaker.
4. Spectral features are extracted from the speech frames. A reduced spectral representation is produced by passing the speech frame through a pseudo filter bank designed to match the frequency sensitivity of the ear. This type of filter bank is called a **mel-scale filter bank** and is used extensively for speech-recognition tasks.
5. Passing the speech frame through a pseudo filter produces a spectral representation consisting of log magnitude values from the speech spectrum sampled at linear 100-Hz spacing below 1000 Hz and sampled at a logarithmic spacing above 1000 Hz.
6. For 4-kHz bandwidth speech e.g., telephone quality speech, this reduced spectral representation has twenty-four log magnitude spectrum samples.
7. The log magnitude spectral representation is then inverse Fourier transformed to produce the final representation, called cepstral coefficients.
8. The last transform is used to de-correlate the log magnitude spectrum samples. We base the decision to use mel-scale cepstral coefficients on good performance in other speech-recognition tasks and a study that com- 20-msec window pares several standard spectral features for speaker identification
9. The sequence of spectral feature vectors extracted from the speech signal is denoted
$X_i = [X_1\ X_2\ X_3 \ldots\ldots X_t]$ which is the sequence feature vectors belonging to speaker i
where the set of cepstral coefficients extracted from a speech frame are collectively represented as a *D*-dimensional feature vector *Xt'* and where *t* is the sequence index and *T is* the number of feature vectors.
10. The set of speakers is given by
$S = [S_1\ S_2\ S_3 \ldots S_N]$
i = 1, 2, 3 … N and N which is the total number of speakers
11. The spectral feature vectors undergo channel compensation to remove the effects of transmission degradation. Caused by noise and spectral distortion, this degradation is introduced when speech travels through communication channels like telephone or cellular phone networks.
12. The resulting spectral sequence representation is the starting point for almost all speech-related tasks, including speech recognition and language identification

A. Statistical Speaker Model

From a given speaker, S, the probability of observing Z is given by equation 3

$$p(Z\,|\,S) = \sum_{K=i}^{K} w_{S,K} N(Z;\mu_{S,K},\Sigma_{S,K}) \qquad (3)$$

$$N(Z;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n\,|\Sigma|}} EXP(-\tfrac{1}{2}(Z-\mu)\Sigma^{-1}(Z-\mu)) \qquad (4)$$

Where $w_{s,k}$ = mixture weight

$\mu_{s,k}$ = mean matrix

$\sum_{S,K}$ = covariance matrix

n = size of Z

$\sum$ = diagonal covariance matrices

The mean vector represents the expected spectral feature vector from the state, and the covariance matrix represents the correlations and variability of spectral features within the state.

In addition to the feature-vector production being a state-dependent random source, the process governing what state the speaker model occupies at any time is modeled as a random process. The above definition of the statistical speaker model is known more formally as an ergodic hidden Markov model (HMM). HMMs have a rich theoretical foundation and have been extensively applied to a wide variety of statistical pattern-recognition tasks in speech processing and elsewhere. The main motivation for using HMMs in speech-recognition tasks is that they provide a structured, flexible, computationally tractable model describing a complex statistical process. Hidden state, weighted by the probability of being in each state. With this summed probability we can produce a quantitative value, or score, for the likelihood that an unknown feature vector was generated by a particular GMM speaker model.

$$\{p1,.....p_N\}, where \sum_{i=1}^{N} p_i = 1 \qquad (5)$$

$$p(X \mid \lambda) = \sum_{i=1}^{N} p_i b_i(X) \qquad (6)$$

$$\lambda = (p, \mu, \Sigma),\ for\imath = 1, 2, ...N \qquad (7)$$

$$S = \arg\max_{1 \le S \le N} \sum_{t=1}^{N} \log p(X_N \mid \lambda_S) \qquad (8)$$

$$\Lambda(X) = \log p(X \mid \lambda_C) - \log p(X \mid \lambda_{\bar{C}}) \qquad (9)$$

$$\log p(X \mid \lambda_{\bar{C}}) = \log\left\{\frac{1}{B}\sum_{b=1}^{B} p(X \mid \lambda_b)\ \right\} \qquad (10)$$

where
X_t =feature vector
Λ(X) =likehood ratio
For Λ(X)≥ θ, accept the claimer
Λ(X) <θ, reject the claimer
θ = decision threshold
λ = sum of the probability that X_t was generated from the hidden state
N = no of feature vectors
P(X| λ_c) = likehood that the utterance belongs to the claimer
P(X| λ_c^-)) = likehood that the utterance does not belongs to the claimer

VI. DATA ANALYSIS AND DISCUSSION

The YOHO database was designed to support text-dependent speaker-verification research such as is used in secure-access technology. It has a well-defined train and test scenario in which each speaker has four enrollment sessions when he/she is prompted to read a series of twenty-four combination-lock phrases. Each phrase is a sequence of three two-digit numbers. There are ten verification trials per speaker, consisting of four phrases per trial. The vocabulary consists of fifty-six two-digit numbers ranging from 21 to 97. The speech was collected in an office in a school environment. Thus the speech has a telephone bandwidth of 3.8 kHz, but no telephone transmission degradations.

The YOHO database is different from the above text-independent, telephone-speech databases, which allows us to demonstrate how the GMM verification system, although designed for text-independent operation, can also perform well under the vocabulary- dependent constraints of this application.

A. Session Detail

There are three major Nigerian languages; Hausa, Igbo and Yoruba. Although, three are over 250 variety of Nigerian languages with different mother tongue, dialect and phonetics. The sessions are divided into English sessions, and mother tongue sessions. For male, three sessions are for English session and two sessions are for mother tongue session while that of female, two sessions are for English session and a session for mother tongue session. These are the training sessions but for testing session, three sessions are used for male and two sessions are used for female.

Table 1: Number of Sessions and Population

Language	**Male**		**Female**	
	#	**# sessions**	**#**	**# session**
Igbo	**20**	**8**	**15**	**5**
Hausa	**20**	**8**	**5**	**5**
Yoruba	**30**	**8**	**30**	**5**

Twenty-20msec segments, or frames, of speech are passed through a speech activity detector, which discards silence and noise frames that reflect the environment rather than the speaker. Spectral analysis extracts spectral features from the speech frames.

Channel compensation removes the effects of transmission degradation from the resulting spectral representations. The success of statistical pattern-recognition approaches for a wide variety of speech tasks; we adapt a statistical formulation for such a speaker model.

In the statistical speaker model, we treat the speaker as a random source producing the observed feature vectors. Within the random speaker source, there are hidden states corresponding to characteristic vocal-tract configurations.

When the random source is in a particular state, it produces spectral feature vectors from that particular vocal-tract configuration. The states are called hidden because we can observe only the spectral feature vectors produced, not the underlying states that produced them.

For the speaker recognition analysis of the voice samples, the wavelet toolbox of MATLAB® 2007 was used. Applying the approach on text independent YOHO testing data, the YOHO database shows that low error rates are possible for a secure-access verification application even with a text-independent verification system. An overall EER of 0.51% and a false rejection rate of 0.65% at a 0.1% false-acceptance rate were obtained. The constrained vocabulary along with the good-quality speech allowed the model to focus on the sounds that characterize a person's voice without extraneous channel variability. The experimental results show, speaker-recognition performance is indeed at a usable level for particular tasks such as access-control authentication.

V. CONCLUSION

With the GMM as the basic speaker representation, this model can be applied to specific speaker-recognition tasks of identification and verification. The identification system is a straightforward maximum likelihood classifier. Applying the approach on text independent YOHO testing data, the YOHO database shows that low error rates are possible for a secure-access verification application even with a text-independent verification system. An overall EER of 0.51% and a false rejection rate of 0.65% at a 0.1% false-acceptance rate were obtained. The constrained vocabulary along with the good-quality speech allowed the model to focus on the sounds that characterize a person's voice without extraneous channel variability. The experimental results show, speaker-recognition performance is indeed at a usable level for particular tasks such as access-control authentication. The major limiting factor under less controlled situations is the lack of robustness to transmission degradations, such as noise and microphone variability. Large efforts are under way to address these limitations, exploring areas such as understanding and modeling the effects of degradations on spectral features, applying more sophisticated channel compensation techniques, and searching for features more immune to channel degradations. We could achieve Error Rate (ER) of 0.96%. We also conducted the same experiment in identification section, we achieved 0.28% identification error rate. We achieve 0.7% identification error rate using 4.45% of the speaker utterance only.

REFERENCES

i. *A. K. Jain and A. A. Ross, Handbook of Biometric, A. K. Jain, P. Flynn and A. A. Ross, Eds., Springer US, 2008.*

ii. *R. Bolle and S. Pankanti, Biometrics: Personal Identification in Networked Society, K. A. Jain, Ed., Norwel, MA: Kluwer Academic Publisher, 1998.*

iii. *F. O. Ehiagwina and L. O. Afolabi, "Managing Insecurity with Biometric Engineering: An Overview of the Nigerian Experience," in Proc. of the Int. Academic Conf. for Sub-Sahara African Transformation & Development, Ilorin, 2015.*

iv. *L. O. Afolabi, J. O. Azanubi, N. A. Iromini and F. O. Ehiagwina, "Overview of Multimodal Biometric System to solve National Issues: INEC, Nigeria as Case Study," in 4th National Conference: School of Engineering, Federal Polytechnic, Offa, Offa, 2015.*

v. *K. Delac and M. Grgic, "A Survey of Biometric Recognition Methods," in 46th International Symposium Electronics in Marine, ELMAR-2004, Zadar, Croatia, 2004.*

vi. *V. Tiwari, "MFCC and its Applications in Speaker Recognition," International Journal on Emerging Technologies, vol. I, no. 1, pp. 19-22, 2010.*

vii. *F. McGehee, "The Reliability of the Identification of the Human Voice," Journal of General Psychology, vol. XXVII, no. 2, pp. 249-271, 1937.*

viii. *W. D. Voiers, "Perceptual Bases of Speaker Identity," Journal of Acoustic Society of America, vol. XXXVI, no. 1065, 1964.*

ix. *J. B. Allen, "How do Human Process and Recognize Speech," IEEE Transactions on Speech and Audio Processing, vol. II, no. 4, pp. 567-577, 1999.*

x. *D. A. Reynolds, "Automatic Speaker Recognition: Current Approaches and Future Trends," in MIT Lincoln Labotatory, ICASSP (2001), MA, USA, 2001.*

xi. *D. A. Reynolds, "Overview of Automatic Speaker Recognition," in JHU 2008 Workshop Summer School, MA, USA, 2008.*

xii. *L. O. Afolabi, T. O. Adewunmi, O. Olumbe-Salau and F. O. Ehiagwina, "Using Gaussian Mixture Model (GMM) Technique for Automatic Speaker Recognition," in 4th National Conference: School of Engineering,*

xiii. *G. Fant, Acoustic Theory of Speech Production, The Hague: Mouton, 1960.*

xiv. *A. S. Charles and C. Stephen, Securing Biometric Application, Springer International Publishing, 2008.*

xv. *L. Bjomn, S. Jhan, B. Peter, D. Hassan and G. Svante, "The Gunnar Fant Legacy in the Study of Vocal Acoustic," in 10th French Congress on Acoustic , Lyon, 2010.*

xvi. *N. Amy and A. P. Hemant, Forensic Speaker Recognition: Law Enforcement and Counter-terrorism, Springer International Publishing, 2011, pp. 1-186.*

xvii. *H. Hermansky, "Perceptual Linear Prediction," Journal of Acoustic Society of America, pp. 1738-1752, 1990.*

xviii. *H. Hermansky, "RASTA-PLP Speech Analysis Technique," in ICASSP-9 Proceedings, 1992.*

xix. *D. A. Reynolds, "Experimental Evaluation of Features for Robust Speaker Identification," IEEE Trans. Speech Audio Process, vol. II, no. 4, pp. 639-643, 1994.*

xx. *L. G. Bahler, J. E. Porter and A. L. Higgins, "Improve Voice Identification Using Nearest-Neighbor Distance Measure," in International Conference on Acoustics, Speech, and Signal Processing, Adelaide, 1994.*

xxi. *National Science and Technology Council, "Speaker Recognition," National Science and Technology Council.*

xxii. *D. J. Woodward, M. N. Orlans and T. P. Higgins, Biometrics, New York: McGraw Hill Osborne, 2003.*

xxiii. *NIST, "NIST Speaker Recognition Evaluation," [Online]. Available: http:www.nist.gov/speech/tests/spk/index.htm. [Accessed 7 August 2015].*

xxiv. *H. Hermansky, "Human Speech Perception: Some Lessons from Automatic Speech Recognition," in Text, Speech and Dialogue, 4th International Conference, 2001.*

xxv. *R. Tong, B. Ma, D. Zhu, H. Li and E. S. Chng, "Integrating Acoustic, Prosodic and Phonotactic Features for Spoken Language Identification," in Acoustic, Speech and Signal Processing, 2006. ICASSP 2006*

xxvi. *S. V. Gite and J. V. Shinde, "A Review on Robust Language Identification," International Journey of Computer Technology and Applications, pp. 21-24, 2016.*

xxvii. *T.-w. Lee and S. S. Lewicski, "The Generalized Gaussian Mixture Model Using ICA," 2000.*

xxviii. *P. KaewTraKulPong and R. Bowden, "An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection," in 2nd European Workshop on Advanced Video Based Surveillance Systems,*

xxix. *P. A. Torres-Carrasquillo, D. . A. Reynolds and J. R. Deller, Jr, "Language Identification using Gaussian Mixture Model Tokenization," 2002.*

xxx. *Z.-K. Huang and K.-W. Chau, "A New Image Thresholding Method Based on Gaussian Mixture Model," Applied Mathematics and Computation, vol. CCV, no. 2, pp. 899-907, 2008.*

xxxi. *R. Singh, B. C. Pal and R. A. Jabr, "Statistical Representation of Distribution System Loads Using Gaussian Mixture Model," IEEE Transactions on Power Systems, vol. XXV, no. 2, pp. 29-37, 2010.*

xxxii. *D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz and S. Thomas, "The subspace Gaussian mixture model—A structured model for* xxxiii. *A. Drygailo and M. El-Maliki, "Speaker Verification in Noisy Environments with Combined Spectral Subtraction and Missing Feature Theory," in Acoustic, Speech and Signal Processing, 1998. ICASSP 1998*

xxxiv. *J. P. Campbell and D. A. Reynolds, "Corpora for the Evaluation Of Speaker Recognition Systems," in Acoustic, Speech and Signal Processing, 1999. ICASSP 1999 Proceedings. 1999 IEEE International Conference on,,*

xxxv. *A. Arvaniti and M. Baltazani, "Greek ToBI: A System for the Annotation of Greek Speech Corpora," in Second International Conference on Language Resources and Evaluation, Athens, 2000.*

xxxvi. *J. Harrington, The Phonetic Analysis of Speech Corpora, Wiley-Blackwell, 2010, pp. 1-424.*

xxxvii. *D. . A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing, vol. X, pp. 19-41, 2000.*

xxviii. *B. Bozkurt, O. Ozturk and T. Dutoit, "Text Design for TTS Speech Corpus Building Using a Modified Greedy Selection," in EUROSPEECH 2003, Geneva, 2003.*

xxxix. *S. Narayanan, E. Bresch, P. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, M. Proctor, V. Ramanarayanan and Y. Zhu, "A Multimodal Real-Time MRI Articulatory Corpus for Speech Research," in Interspeech 2011, 2011.*

xl. *A. Park and T. J. Hazen, "ASR Dependent Techniques for Speaker Identification," in Proceedings of the 7th International Conference on Spoken Language Processing, 2002, Denver, Colorado, 2002.*

xli. *L. Sang, Z. Wu, Y. Yang and W. Zhang, "Automatic Speaker Recognition Using Dynamic Bayesian Network," in Acoustic, Speech and Signal Processing, 2003. ICASSP 2003 Proceedings. 2003 IEEE International*

xlii. *J. Bai, R. Zheng, B. Xu and S. Zhang, "Robust Speaker Recognition Integrating Pitch and Wiener Filter," in Chinese Spoken Language Processing, 2004 International Symposium on, 2004.*