Speaker forensic identification using joint factor analysis and i-vector

R J Rouf, and D Arifianto*
Engineering Physics Department, Institut Teknologi Sepuluh Nopember, Indonesia

*dhany@ep.its.ac.id

Abstract. Speaker Forensic is a process to determine the identity between a person's voice (known speaker) and the investigated voice (suspect speaker). To improve accuracy in speaker forensic analysis is used a combination of 2 forensic methods, joint factor analysis and i-vector methods. The forensic approach by adding noise signal with SNR value as a representation of a tapping situation to measure speaker identification performance. Classification on verification using i-vector is done by comparing i-vector model tests and targets. Both models calculated vector similarities using cosine similarity score. Verification is done with compare the same speaker and verification between different speakers The testing process on the program performance is indicated by an equal rate error value. While the system's sensibility is indicated by the threshold value. The results showed that the EER value of the Graz dataset (1.87%) compared to the Indonesia dataset (10%).

1. Introduction
One of the pieces of evidence that are widely used in various types of cases is in the form of audio recordings. Generally, the recording will be a conversation that is evidence of a suspected act. Most recordings are in the form of digital audio, then efforts to analyze audio recordings are carried out using the speaker forensic approach. Speaker forensic is a process to determine the identity between a person's voice (known speaker) and suspect speaker[1]. If the results of the speaker forensic analysis are used as the main evidence to decide a case in court, then it takes a speaker forensic system that accurately with the smallest error reading possible.

The challenges of the speaker recognition research community over the past few years have been how to address session variability issues and channel discrepancies caused by factors such as speaker emotional state, environmental conditions, recording devices, different transmission lines, etc [2]. These factors drastically decrease the classical system’s performance based on the Gaussian Mixture Model-Universal Background Model (GMM-UBM) paradigm [3]. Despite the fact that the GMM-UBM system’s reliability for session variability can be improved with various features, models, and score compensation techniques, the most successful achievements resulting in the best performing speaker verification system are based on learning the variability of this session. In this context, Joint Factor Analysis (JFA)[4] and i-vector Total Variability modeling[5] have gained prevalence across almost all advanced independent text-speaker verification systems. Recently, Mandasari et al [6] studied the effects of speech duration on i-vector system score calibration to evaluate evidence’s strength using the NIST SRE-2010. In Dehak et al [7], carried out an experiment which proves that channel factors estimated using JFA, which are supposed to model only channel effects, also contain information about
speakers. Based on this, Dehak proposed a new speaker verification system based on factor analysis as a feature extractor[8]. The factor analysis is used to define a new low-dimensional space named total variability space. In this new space, a given speech utterance is represented by a new vector named total factors (we also refer to this vector as “i-vector” in this paper)[5]. While i-vectors were originally proposed for speaker verification, they have been applied to many problems, like language recognition, speaker diarization, emotion recognition, age estimation, and anti-spoofing[9].

In the speaker forensic analysis, generally, the number of data is limited and urgently needed. In general, voice recognition systems use UBM-GMM to obtain general characteristics of speakers. One of the main difficulties of the GMM-UBM system is that it involves intercession variability[10] [11]. Joint factor analysis (JFA) is proposed to compensate for this variability by modeling inter-speaker variability separately and channel or session variability and ignoring channel factors. However, it was found that there is information about speakers on channel factors in JFA[7], so the total variability of space is a development of the classic Joint Factor Analysis used to compensate for information about the speaker within the channel factor of the JFA by knowing the total variability of its space. Intersession variability was then compensated for by using backend procedures, such as linear discriminant analysis (LDA) and within-class covariance normalization (WCNN)[5]. While i-vector is used to extract total factors or hidden variables using the Baum-Welch statistic to find unknown parameters[11]. The use of the cosine kernel as a decision score for speaker verification makes the process faster and less complex than other JFA scoring methods[12]. Therefore, this research aims to create a speaker forensic identification system with the Joint Factor Analysis and i-vector methods.

2. Signal in Vector

2.1. i-vector Extraction

i-vector models for the sound of the voice can be obtained using the following equation:

\[
w = (I + T^c \Sigma^{-1} N(u)T)^{-1} T^c \Sigma^{-1} \hat{F}(u)
\]

(1)

Where, \(N(u)\) is defined as the diagonal matrix of CF \(\times\) CF dimensions whose have diagonal block is \(N_c I\) ( \(c = 1, ..., C\)). \(\hat{F}(u)\) is a supervector CF \(\times\) 1 dimension obtained by combining all the first order Baum-Welch \(\hat{F}\) statistics \(c\) for speech given \(u\). \(\Sigma\) is a diagonal covarian matrix of CF \(\times\) CF dimensions estimated during training analysis [4] and the model is residual variability that is not captured by the total variability of the \(T\) matrix[5].

2.2. Performance Indicators System

![Figure 1. Distribution of target and non-target scores[13]](image)

The speech recognition system’s performance indicator can be seen from the target score distribution when the Known and Unknown speeches are the same, and the distribution of non-target scores when known and unknown speeches are different shows by Figure 1. To measure the performance of speech
recognition system used the Equal Error Rate (EER) value. The speech recognition system is a binary-classifier system or binary class separator, where there are only two kinds of outputs or classes. These classes are target and non-target score classes. There are two possible errors in other binary class separator systems False Rejection Rate (FRR) and False Alarm Rate (FAR) errors. FRR indicates the probability of misclassification of the target score to be non-target, while FAR indicates the probability of misclassification of the non-target score being the target. The EER value is obtained when the FRR and FAR values are the same. The EER value is a commonly used indicator of the performance of the auto-speech recognition system. The EER value indicates the performance of discrimination between the target and non-target scores. The smaller EER value indicates the better the performance of discrimination a system [13].

3. Methodology

3.1. Data Collection
This research used secondary data in the form of voice data .wav format derived from graz dataset [14] and Bahasa Indonesia dataset [15]. In graz dataset there are 2 types of data, namely clean dataset and LAR dataset (dataset with added noise reverberant) in 20 speakers (10 female and 10 male). While the Indonesian dataset contains clean data with 8 speakers (4 female and 4 male).

3.2. Feature Extraction
Feature extraction is the retrieval of MFCC, delta MFCC, and delta-delta MFCC features that contain information in the form of a character feature of the voice-to-voice signal to be analyzed. The feature extraction stage is done by extracting the feature using 25 ms hanning Windows analysis with 10 ms overlap and window length 9. Feature extraction is performed to take 20 dimension features MFCC on each sound audio signal. Coupled with the coefficient of delta and delta-delta MFCC, 60 dimension features are formed.

3.3. Training UBM-GMM
Systems that are training with TIMIT datasets usually contain about 2048 ubm components. However in this case it only uses 256 components. The first step is the expectation step of the EM algorithm, where adequate statistical estimates of speaker training data are calculated for each mix in UBM. P (x|l) is a posterior likelihood function that can be calculated using following equation. [3]

\[ p(x|\lambda) = \sum_{i=1}^{M} w_i p_i(x) \] (2)

Equation 2 is used to adjust the mixture density for the D-dimensional vector (x) feature. While gaussian M density is unimodal, pi (x), each parameter with vector D x 1 average (\mu_i) and matrix covarianis D x D (\Sigma_i) can use equation 3 [3].

\[ p_i(x) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu_i)'(\Sigma_i)^{-1}(x - \mu_i) \right\} \] (3)

Calculating the probability of posterior normalization:

\[ \log p(X|\lambda) = \sum_{t=1}^{T} \log p(x_t|\lambda) \] (4)

While the second step of the EM algorithm is maximization is used to adapt the new sufficient statistic estimates then combined with old sufficient statistics from UBM mixture parameters using data-dependent mixing coefficient. UBM and vector training from the hypothesized, X ={x1, .., xT},
Specifies the probabilistic alignment of the training vector into the mixed component i.e. for the mixture in UBM, can be calculated with equation 5.

\[
Pr(\hat{i}|x_t) = \frac{w_i p_i(x_t)}{\sum_{j=1}^{M} w_j p_j(x_t)}
\]  

(5)

Other parameters produced are ubm.mu (component means) and ubm.sigma (component covariance matrices).

3.4. Statistic Baum-Welch

Baum-Welch statistics are used to compensate for these hidden variables. Baum-Welch statistics are \( N \) (zero order) and \( F \) (first order) used in EM algorithms, can be calculated using equations 6 and 7.

\[
N_c = \sum_{t=1}^{L} P(c|y_t, \Omega) = \sum_{t} \gamma_t(c)
\]  

(6)

\[
F_c = \sum_{t=1}^{L} P(c|y_t, \Omega)y_t = \sum_{t} \gamma_t(c) y_t
\]  

(7)

Gamma \( \gamma_t(c) \) generated from exp (loglikehood – loglikehoodSum), loglikehood and loglikelihoodSum are variables resulting from the calculation of posterior log likelihood in ubm expectation steps. \( N_c \) is the sum of gamma, while \( F_c \) is the result of gamma multiplication and the result of extracting features from audio files.

3.5. Total Space Variability

The total variability space (T) contained in the following equation:

\[
M = m + Tw
\]  

(8)

I-vector extraction is characterized by \( m \) as an average supervector (ubm), \( T \) is a low rank matrix of total variability, \( \Sigma \) is a diagonal covarian matrix. Here are the steps to find total variability space:

3.5.1. Calculates the posterior distribution of the hidden variable \( l(u)w(u) \). with the following equations:

\[
l(u) = I + T^\top \Sigma^{-1} N(u)\Sigma
\]  

(9)

Where, \( N(u) \) is the zero order of baum-welch statistic, \( I \) is the matrix of identity, \( T \) and \( \Sigma \) total variability and covarian matrix or ubm.sigma.

3.5.2. Accumulate statistics on all speakers. Expressed with the following equations:

\[
E[w(u)] = l^{-1}(u)T^\top \Sigma^{-1} \tilde{F}(u)
\]  

(10)

\[
C = \sum_u \tilde{F}(u)E[w^\top(u)]
\]  

(11)

\[
A_c = \sum_u N_c(u)l^{-1}(u)
\]  

(12)

Where, \( \tilde{F}(u) \) is the centralized first order statistic, \( l^{-1}(u) \) is the kovarian matrix resulting from the matrix inverse, \( T \) and \( \Sigma \) total variability and covarian matrix, \( A_c \) is the accumulation of statistics for the entire speaker \( l(u) \).

3.5.3. Total space variability. Can be expressed through the following equations:

\[
T_c = A_c^{-1}C
\]  

(13)

\[
T = \begin{bmatrix}
T_1 \\
T_c
\end{bmatrix}
\]  

(14)

\( A_c^{-1} \) is the matrix inverse of \( A_c \). Input from the process of total space variability in the form of ubm.sigma \( \Sigma \) from UBM variance in accordance with the following equations:
\[
\sum = N^{-1}\left(\sum_{u} \tilde{S}(u) - \text{diag}(CT^t)\right)
\]  

(15)

Where, \( \tilde{S}(u) \) is the centralized second order statisticbaum-welch, \( c = 1, \ldots, c \) is the number of ubm components, generally in TIMIT data used as much as 1000 but in this program used 256 components.

3.6. i-vector Extraction

After getting the result of total space variability matrix (T), then mathematically i-vector can be calculated using equation 1. To inverse the matrix of calculations in \((1 + T^t\Sigma^{-1}N(u)T)\) the equation used pseudo-inverse in order to inverse the non-square matrix. Then get the tranpose of matrix operation to get the i-vector model for training data.

3.7. Projection Matrix

This stage is used to compensate for intercession in the speaker identification process. In this case used LDA and WCNN as projection matrix on this simulation.

LDA is used to minimize intra-class variance and maximize variance between speakers. It can be calculated as described in [7], equations 16 and 17 below:

\[
S_b = \sum_{s=1}^{S} (\bar{w}_s - \bar{w})(\bar{w}_s - \bar{w})'
\]

(16)

\[
S_w = \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (w_i^s - \bar{w}_s)(w_i^s - \bar{w}_s)'
\]

(17)

With \( \bar{w}_s \) is the average i-vector for each speaker, \( \bar{w} \) is the average i-vector on the outside of the speaker, \( n_s \) is the amount of sound said for each speaker. Thus the equation of 18 eigenvalue for eigenvectors is as follows:

\[
S_b v = \lambda S_w v
\]

(18)

While WCNN is used to scale i-vector space inversely proportional to in-class covariance, so that the direction of high intra-speaker variability is not emphasized on i-vector [5], can be stated in equation 19 below:

\[
W = \frac{1}{s} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (w_i^s - \bar{w}_s)(w_i^s - \bar{w}_s)'
\]

(19)

With \( \bar{w}_s \) is the average i-vector for each speaker, \( n_s \) is the total audio file used for each speaker. So for B use Cholesky decomposition as follows:

\[
W^{-1} = BB'
\]

(20)

Equation 18 is used to find eigenvalue for eigenvectors. The number of vector Eigen’s used is 17. While look for B which is the eigenvalue of the WCNN Matrix can be obtained with the equation of 20 decomposition Cholesky. B is the result of the factorization of Cholesky, in the MATLAB to calculate the factorization can use the function Chol.

3.8. Verification Stage

Assessment technique uses the kernel cosinus value between the target speaker i-vector (wtarget) and the i-vector (wtest) test as a decision score with an equation of 21 [16]:

\[
\text{score}_{(w_{target}, w_{test})} = \frac{(w_{target}, w_{test})}{\|w_{target}\|\|w_{test}\|} 2\Theta
\]

(21)
CSS FRR is the result of the verification stage to the same speakers test. In different speaker tests, the test data is randomly retrieved outside the target speaker data but still within the scope of the speaker id on the target speaker data as many as 100 file for each speaker. CSS FAR is the result of the verification stage to different speakers test.

4. Result and Discussion
Based on the research that has been done obtained the following results:

4.1. Dataset Program
In the simulation, Graz datasets will be conducted intact without the addition of noise. However, for Indonesia datasets will be added interference in the form of babble noise and rain noise as an approach to tapping situations in speaker forensics. Rain noise is obtained from sound recordings with a duration of 30 seconds. Meanwhile, babble noise is a noise disorder created by combining multiple sound signals from several different speakers.

![Figure 2. Sound signal waveform of Indonesia Dataset](image)

From figure 2 can be seen the type of sound dataset and its differences through the waveform plot of the sound signal of each type of sound dataset (clean data, with rain noise and babble noise). This type of dataset intended to test the performance of the program, while also knowing the reliability of some types of data in many conditions.

4.2. Babble Noise Evaluation
The Babble noise created by combining 3 sound files from 3 different speakers. Analysis of the distribution of babble noise statistics is carried out to find out the characteristics of the noise modeling. Figure 3 shows the statistical distribution form of babble noise. Such signals can be categorized as non-Gaussian or unknowable from the probability normality test [17]. The fitting process is done using MATLAB software dfittool and normality test is carried out using the normality test menu (probability plot) on Minitab software.
Based on the results of fittings and normality tests obtained information that the babble noise is a non-Gaussian signal. This is seen in figure 3 which shows the distribution beyond the normal line reinforced by the normality test plot in figure 4. Therefore the babble noise have a high Speech Intelligibility Index where the level of clarity of sound heard by the listener is good in other words all conversational information can be heard well (audible) [18].

4.3. Performance Evaluation

The evaluation performance of each experiment for each dataset can be shown in table 1. EER is obtained from the intersection of FAR and FRR values. While accuracy is a 100% reduction with EER (%).

Table 1. EER Results and Accuracy for Noise Datasets

| Dataset                        | EER (%) / Performance | Accuracy (%) | Threshold / sensitivity |
|--------------------------------|-----------------------|--------------|------------------------|
| Graz (Clean Data)              | 1.87                  | 98.13        | 39.39                  |
| Graz (LAR Data)                | 6.58                  | 93.42        | 69.20                  |
| Indonesia (Clean Data)         | 10.00                 | 90.00        | 58.59                  |
| Indonesia (with Rain Noise)    | 21.25                 | 71.75        | 84.48                  |
| Indonesia (with Babble Noise-SNR 0) | 38.54                | 61.46        | 96.90                  |
| Indonesia (with Babble Noise-SNR 10) | 23.54               | 76.46        | 92.38                  |
| Indonesia (with Babble Noise-SNR 20) | 17.50               | 82.46        | 57.92                  |

Table 1 shows the EER value, accuracy and sensitivity for each type of dataset. The EER value indicates the performance of discrimination between the target and non-target scores. Smaller EER values show the better performance of a system’s discrimination. While the threshold value indicates the sensitivity of the system so that the greater the sensitivity, have safer system. A good system is a system that has a sensitivity more than 50%.

When compared between clean data, Graz’s EER value is 1.87% while the EER value of Indonesian data is 10% showed by figure 5. A significant difference in EER value from the results of the verification test against Graz dataset and Indonesian dataset can be analyzed due to differences in formant structure and F0 (Fundamental Frequency). From the analysis obtained shows that the formant of the Indonesian dataset is clearer, while F0 dataset Indonesia is more glide. The structural difference is one of the internal aspects related to the difference in value obtained in table 1 between Graz and Indonesian dataset.
Another factor is the quality of the two datasets different quality of recording in the form of the condition of the recording process as well as the tools used.

The results of the experiment on Indonesia dataset with babble noise showed that, higher the SNR value then the smaller the test error obtained. Signal to noise ratio (SNR) indicates the ratio of the power comparison between the signal and the given noise. The greater the SNR value, the smaller the power noise compared to the signal, and vice versa. The given SNR variation provides the result data that appears in table 1. Figure 6 (a)-(c) for clean data, (d)-(f) sound data with Babble noise SNR 20, (g)-(i) sound data with Babble noise SNR 10, and (j)-(l) sound data with Babble noise SNR 0. When compared them to clean data, from figure 6 it appears that the smaller SNR value, the more stacked the formant structure. Because this babble noise is non-gaussian signal that all conversational information can be heard well (audible). While resulting the F0 value is different, from the addition of F0 to the new shifting F0. This is what causes error readings generated by babble noise SNR 0 higher.

When mixing Indonesian and Graz datasets into one dataset, it is generated as shown in figure 7 (a) shown by the blue bar chart. The testing scheme uses 8 speakers where each dataset is taken by 4 speakers (2 female and 2 male). In figure 7 shows one of the EER mixed dataset test results is 2.5% on the maximum number of training speakers. The EER value represents the performance of a combined system from the Indonesian and Graz datasets. The EER value generated from the mixed dataset tends to be close to the EER value of the Indonesian dataset which has a higher error than in the Graz EER dataset. Therefore the Indonesian dataset is more dominant to affect the performance of the combined dataset. Figure 7 (b) represents the results of a test to determine the effect of gender on performance, which from the image is obtained for a female mixed dataset and a male mixed dataset to achieve an EER of 2.5%. So the gender does not perfect the performance results of the voice identification system.

![Figure 5. EER Result of (a) Dataset Graz and (b) Dataset Indonesia (Clean Data)](image-url)
5. Conclusion

The system successfully verifies from a mix of multiple speakers with 20 files database for each speaker. The FRR is verification result to the same speakers test, while the FAR is verification result to different speakers test. So that EER performance can be determined when the FRR value is equal to FAR. The EER value indicates the discrimination performance of the verification program. The results showed that the EER value of the Graz dataset (1.87%) compared to the Indonesia dataset (10%). The significant difference in EER value occurs due to differences in formant and F0 (Fundamental Frequency) structures as well as the quality of the two different datasets. Performance is influenced by the number of speakers used at the training stage and is not influenced by gender.
6. References

[1] P. Rose, "Forensic Speaker Identification," 1 ed., USA and Canada: Taylor & Francis, 2002.

[2] Z. M. Boulkenafet, B. H. F, O. Nouali and C. Mohamed, "Forensic evidence reporting using GMM-UBM, JFA and l-vector methods: Application to Algerian Arabic dialect," *International Symposium on Image and Signal Processing and Analysis (ISPA 2013)*, pp. 397-402, 2013.

[3] D. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[4] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modelling with sparse training data," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 3, may 2005

[5] P. Kenny, R. Dehak, P Dumouchel, "Front-End factor analysis for speaker verification," *IEEE Trans. Audio, Speech, and Langauge Processing*, vol. 16, no. 4, pp. 788-798, 2011.

[6] M. Mandasari, "Evaluation of ivector speaker recognition systems for forensic application," *INTERSPEECH’11*, pp. 21-24, 2011.

[7] N. Dehak, "Discriminative and Generative Approches for Long- and Short-Term Speaker Characteristics Modeling: Application to Speaker Verification," Ph.D. thesis, Ecole de Technologie Superieure, Montreal, 2009.

[8] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support Vector Machines versus Fast Scoring in the Low Dimensional Total Variability Space for Speaker Verification," in *Interspeech*, Brighton, 2009.

[9] Verma, Pulkit, and P. K. Das. "I-Vectors in Speech Processing Applications: A Survey." *International Journal of Speech Technology*, vol. 18, no. 4, pp. 529–46, 2015.

[10] P. Kenny, et al. "Joint Factor Analysis versus Eigenchannels in Speaker Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–47, May 2007

[11] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A Study of Interspeaker Variability in Speaker Verification," *IEEE Transaction on Audio, Speech and Language*, vol. 16, no. 5, pp. 980–988, july 2008

[12] O. Glembek, L. Burget, N. Brummer, and P. Kenny, "Comparaison of Scoring Methods used in Speaker Recognition with Joint Factor Analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, April 2009

[13] M. I. Mandasari, A. D. Firmanto and F. Fathurrahman, "Kalibrasi Rasio kemungkinan pada Sistem Rekognisi Penguap Otomatis untuk Aplikasi Forensik di Indonesia," *Jurnal Linguistik Komputasional (JLK)*, vol. 2, no. 2, 2009.

[14] Pernkopf, F., Graz University of Technology (PTDB-TUG), *Signal Processing and Speech Communication Laboratory*:[Online].Available at: https://www.spse.tugraz.at/databases-and-tools/ptdb-tug-pitch-tracking-database-from-graz-university-of-technology.html. [Accessed 12 December 2019].

[15] Anggraini, E. "On Development Of Indonesian Natural Speech Synthesis Based On Hidden Markov Model HMM". *Undergraduate Thesis of Physics Engineering*, 2013.

[16] N. Dehak, R. Dehak, R. James, Glass, D.A Reynolds and P. Kenny. "Cosine Similarity Scoring without Score Normalization Techniques." *Odyssey*, 2010.

[17] Nadiroh, A. "Efektivitas Babble Speech Masker Terhadap Just Noticeable Difference Untuk Peningkatan Privasi Pada Rumah Sakit Dengan Tipologi Perkantoran Tapak Terbuka," Thesis Report, Eng. Physics Dept., Institut Teknologi Sepuluh Nopember, Indonesia, 2019.

[18] B. W. Hornsby, "The Speech Intelligibility Index: What is it and what's it good for?", *The Hearing Journal*, pp. 10-17, 2004.