Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

Text-Dependent Speaker Verification System in VHF Communication Channel

Chang Huai You, Kong Aik Lee, Bin Ma, Haizhou Li

Institute for Infocomm Research (I^2R), A*STAR, Singapore 138632

{ echyou, kalee, mabin, hli }@i2r.a-star.edu.sg

10.21437/Odyssey.2014-33

Abstract

Text-independent speaker verification can reach high accuracy provided that there are sufficient amount of training and test speech utterances. Gaussian mixture model - universal background model (GMM-UBM), joint factor analysis (JFA) and identity-vector (i-vector) represent the dominant techniques used in this area in view of their superior performance. However, their accuracies drop significantly when the duration of speech utterances are much constrained. In many realistic voice biometric application, the speech duration is required to be quite short, which leads to low accuracy. One solution is to use pass-phrases in place of the uncertain contents. In contrast with text-independent system, this kind of text-dependent speaker verification can achieve higher accuracy even when the speech is short. In this paper, we conduct a study on the application of the pass-phrase based speaker modeling and recognition where the speech signal is obtained through VHF (Very High Frequency) communication channel. We attempt to evaluate the effectiveness of the GMM-UBM, JFA, i-vector methods and their fusion system on this text-dependent speaker verification platform. Our primary target is to achieve equal error rate (EER) of 10∼15% under adverse condition using about 3 seconds of speech sample.

1. Introduction

In more recent years, GMM-based systems have been applied successful in speaker recognition field [1, 2, 3]. Techniques such as GMM-UBM, JFA and i-vector achieve high recognition accuracy given sufficiently long segment of speech for text-independent speaker verification. However, the performance usually drops rapidly when they are applied for short duration speech. In contrast, text-dependent speaker verification can achieve considerable high accuracy when the speech is short [4].

In this paper, we advocate the use of pass-phrase based speaker verification to provide an automatic and reliable authentication of far-end speaker through VHF communication channel. Being different from mobile cellular phone communication, the VHF talking channel in our study is of much noisier interference and high varying channel effects. In this regard, we aim at developing a voice biometric technology for its robustness against the distortion originated from VHF communication channel. The ultimate goal is to benchmark state-of-the-art technology and to show whether current technology is ready for deployment using speech samples transmitted through the VHF channel. Our target is to achieve over 85∼90% of accuracy within 3 seconds of speech.

In particular, the speaker verification system is designed to authenticate the identity of the shipmaster using his name and certificate identity (ID). For this to be possible, the shipmaster has to enroll his voice to the system by pronouncing his 'name' and 'ID' multiple time (typically three repetitions). We apply an expert system contributed from three GMM-series techniques for the voice biometric system. They are GMM-UBM, JFA and i-vector.

The GMM-UBM technique has shown reliable performance for text-independent speaker recognition [5, 6, 7]. A GMM carries rich amount of information including speaker information, speech contents, channel and emotion from its corresponding utterance [8, 9, 10]. JFA approach [11] is effective due to its efficient modeling of speaker factors [12] and channel factors [13], whereby a GMM-supervector is viewed as a combination of speaker and channel specific supervectors. It compensates for the channel variation through eigenchannel modeling and emphasizes the speaker-dependent component through eigenvoice modeling. More recently, the i-vector technique, originated from JFA, brings a new height to speaker recognition and has become the most popular [14, 15]. The i-vector extractor converts a sequence of features into a low-dimensional vector in the total variability space, by which speech segment of variable length can be represented as fixed-length vector. In this regard, linear discriminant analysis (LDA) [16], probabilistic LDA (PLDA) [17, 18], and the heavy-tailed PLDA [19, 20] are useful for i-vector system.

In this paper, we introduce the features of VHF communication and develop a strategy to overcome the problem of VHF channel through both feature extraction stage and model training design. Against the very short speech duration in verification, we give a pass-phrase based modeling scheme. It is observed that the score distribution of the imposter trial with correct pass phrase is the closest to that of the genuine trial with correct pass phrase. It implies that the pass phrase is more informatively important than who the speaker is. Finally, we report the performance of GMM-UBM, JFA and i-Vector systems and show how effective their fusion is on a pass-phrase based speaker verification platform.

In the remainder of the paper, the VHF communication channel effect for speaker verification is introduced and the pass-phrase modeling and recognition structure are developed in Section 2. Different speaker verification systems are briefed in Section 3. The task platform for text-dependent speaker verification is described and the performance measure is reported in section 4. The conclusion is given in Section 5.

2. VHF Communication Channel and Modeling Strategy For Speaker Verification

2.1. VHF Communication Channel

2.1.1. About VHF

The VHF range of the radio spectrum is the band extending from 30 MHz to 300 MHz, while the ITU (International Telecommunication Union) defines the marine VHF band as

the radio frequency range between 156.0 and 162.025 MHz. The wavelengths corresponding to these limit frequencies are 10 meters and 1 meter.

In the VHF band, electromagnetic fields are affected by the earth's ionosphere and troposphere. Ionospheric propagation occurs regularly in the lower part of the VHF spectrum, mostly at frequencies below 70 MHz. With ionospheric propagation, the communication range can sometimes extend over the entire surface of the earth. The troposphere can cause bending, ducting, and scattering, however it can still extend the range of communication significantly beyond the visual horizon. In other words, VHF can be reflected, reduced or even stopped by other objects. It can travel between 35-50 miles offshore. The higher the VHF power is the further the range it travels. Sufficient power can improve the quality of transmitted signal and also overcome some obstacles.

The marine VHF uses frequency modulation to convey voice. When the carrier wave is modulated, sideband signals are produced that deviate above and below the carrier frequency. To prevent signals from interfering with signals on adjacent channels, the spacing between channels is set to roughly twice the modulated signal width, or 50-kHz for a modulation bandwidth of 25-kHz of each carrier frequency.

2.1.2. Why VHF

In technical terms VHF is similar to the way that commercial radio stations transmit. Its equipment is relatively simple, and can therefore be compact and low cost. The propagation distance of VHF is limited in a small area, it may not be interfered from other VHF users who use the same frequency band in a far area.

In one side, the VHF band is popular for mobile two-way radio communication with boats. There are a number of communication devices, including cellphones and more sophisticated communication devices. However for the majority of boat owners, a VHF is about as good as you need. Cellphones coverage is limited to areas of higher population density, while VHF Marine coverage is extensive so a call will likely be heard by someone, whether coastguard or a private listening station. In another side, the propagation characteristics of VHF are suitable for short-distance terrestrial communication, with a range somewhat farther than line-of-sight from the transmitter.

2.1.3. The Problem

The development data considered in this paper was collected with the participation of twenty two port inspectors (PIs) of the maritime and Port Authority of Singapore (MPA) [21]. The communication between ship and the Port Operations Control Centre (POCC) is typically carried out with the use of VHF radio. In this regard, the voice recordings were collected by having the PIs pronouncing a list of sentences though the marine VHF radio from a PI ship located at Singapore strait to the POCC located at the west coast of Singapore. In this paper, we used both channels 6 and 7 (corresponding to central frequencies of 156.300MHz and 156.350MHz respectively) of the VHF radio meant for ship-to-ship and ship-to-shore communication.

On the contrary with the National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation (SRE) [22] speech database, the main characteristics of the VHF speech data used in our system are: 1) very short speech duration ($1 \sim 3$ seconds); 2) open channel mismatch problem; 3) strong noisy environment; 4) high distortion of communication channel. Fig. 1 shows the spectrogram of a speech signal

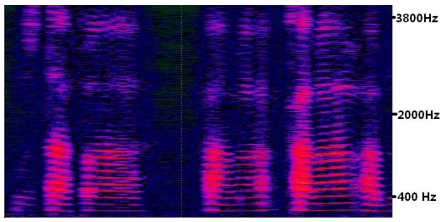

Figure 1: Speech spectrogram recorded using a close-talk microphone.

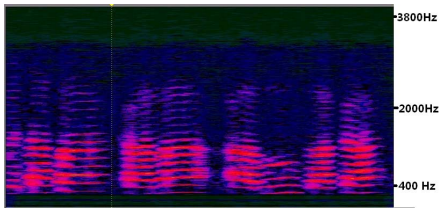

Figure 2: Speech spectrogram received through VHF channel from a ship.

recorded using a close-talk microphone in office environment while Fig. 2 the spectrogram of the speech received through VHF communication channel between on board a ship and the control center. In the task, the VHF speech recorded in office is used for enrollment, and the VHF speech from on board used for test. The major factor that affects the performance is the channel distortion when speech signals were wirelessly transmitted through the VHF channel. The above limitations lead to low performance of the conventional speaker verification algorithm applied to the VHF platform. For example, the JFA algorithm is capable of 0.8539% of EER for the NIST SRE2012 Eval task [23], however it drops down to 18% of EER for the VHF platform.

In contrast with the telephone communication speech signal, VHF signal has narrow bandwidth limitation. The VHF channel restricts the bandwidth of the baseband speech signal to be much lower than 3.2 kHz. This is much lower than the 3.4kHz telephone bandwidth. The information contained in the lower bandwidth corresponds mostly to the linguistic content of speech (i.e., the words and messages to convey via speech). It has been shown in a number of reports that the higher bandwidth contains useful information linked to the voice characteristic of a person [24]. As such, with this higher bandwidth removed from the VHF channel, the speaker information in the VHF signal is greatly suppressed.

In addition to the reduction in speech bandwidth, Table 1 shows other intrinsic and extrinsic differences between the voice samples of the same speaker when recorded on-board and in-office. On the intrinsic side, it was found that some speakers tend to speak faster when they are recording through the VHF. This is reflected from the amount of on-board data, for example, some speakers took in average 0.91 second per recording compared to average 1.3 seconds per recording. However, further analysis discovers that the same speakers have a faster speaking rate compared to other speakers even for in-office recordings 0.95 and 0.98 seconds per recordings for the speakers, respectively. As such, speaking rate seems not to be a limiting factor here. Another intrinsic factor is the vocal effort. It was found that the speakers tend to raise their voice (toward shouting) for onboard recording. This is common to most speakers. Recent

Table 1: Difference between inoffice and onboard recording

Intrinsic factors	Extrinsic factors
Speaking rate	Channel distortion
Vocal effort	Background noise
	Voice deformation

study shows that speaker characteristics drifted when vocal effort changes from normal to shouting.

On the extrinsic side, the VHF channel degrades the quality of the speech signal. The degradations manifest themselves in an unpredictable manner as voice deformation (change of timbre), additional of background noise correlated to the speech signal, and other forms of channel distortion (e.g., the harmonic distortion due to clipping). Source of distortions: (a) Fading of signal when the distant changes. (b) Electromagnetic distortion from surrounding devices, including the engine.

2.1.4. Solution

Against the problem of the VHF speech database, firstly, we developed a robust voice-activity-detection (VAD) for feature extraction. In particular, we applied spectral subtraction denoise [25] on the raw speech signal, and the VAD is analyzed based on the denoised speech. It is observed that applying denoise only to VAD but not to speech for feature extraction is the best selection in current experiment database [26].

Secondly, we focus on database collection and organization. We simulated the particular application to record the speech database using VHF device. Those data are to involve the particular system parameter training. Through the design of training data assignment, we increase the robustness of the verification system. Against the complexity of the VHF speech signal, we collected many speech data to simulate the enrollment condition and verification condition. Proper design of the database assignment is a key to strengthen the robustness of the speaker verification system against the noise, channel mismatching and signal distortion. The databases include **I2R-2013-Data**, **iPad-Data**, **FourSpks-inoffice**, **Onboard-2010-5spks-shipmaster** and **RSR2015** [27, 28]. The features of the abovementioned databases are listed in Table 2. They are used for UBM training and all other system parameter training including diagonal-matrix, eigen-channel matrix and PLDA etc.

The voice biometric system is a combination of hardware and software: The hardware component consists of a computer and a USB sound card for speech signal acquisition via the VHF handsets. The software component consists of the voice biometric engine, a user management system for enrollment of speakers, and a user interface for operator. An example setup of the prototype is shown in Fig. 3. Here, two Motorola GP328 portable radios are used to set up the VHF communication channel. In the actual operation, the VHF signals will be taken from the marine VHF system. The operational frequency used in our current setup is 170.375 MHz, which is slightly higher than the marine VHF in the range from 156 MHz to 162.025 MHz. In the experiment, the power of portable devices is much low and therefore could only serves shorter distance of VHF transmission.

Table 2: Databases used for system's parameter training

Data	Environment, Recording-Way	Recording Device
I2R-2013-Data	office, (1) VHF with ($<$ 2 meters) close and ($>$ 10 meters, obstacle) far distance, (2) close-speech-direct-record	(1) Walkie-Talkie mic and (2) normal-mic
RSR2015	office, direct-cellular-record	mobile device
iPad-Data	office, direct-cellular-record	iPad device
FourSpks-inoffice	Office, VHF the same as DEV\EVAL set	Walkie-Talkie mic
Onboard-2010-5spks-shipmaster	Onboard, VHF	Walkie-Talkie mic

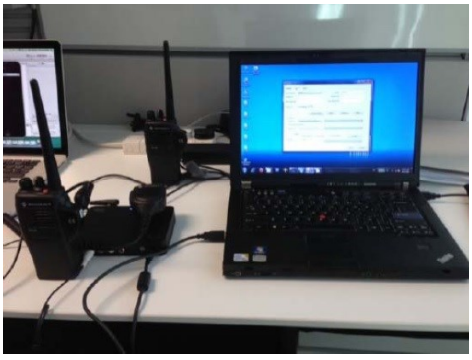

Figure 3: A prototype of the VHF speaker verification system.

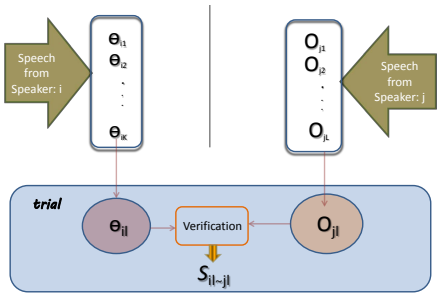

Figure 4: VHF modeling and verification trial. K is the number of pass-phrase models trained for speaker i; L is the number of pass-phrase used for verification for speaker j who want to claim the identity of speaker i.

2.2. Modeling Strategy

Fig. 4 shows the modeling and verification. In a claim, if the maximum score can not be higher than the likelihood threshold, it means the claim is not true. In the pass-phrase based speaker verification system, the different models are trained with corresponding pass-phrases for each target speaker. As a result, there are a series of models $(\Theta_{ik}, k \in \{1, 2, ..., K\})$ for speaker $i, (i \in \{1, 2, ..., I\})$. In verification, the purpose is to verify the identity of speaker j speaking pass-phrase l $(\mathrm{O}_{jl}, l \in \{1, 2, ..., L\})$ whether it is spoken from speaker i or not.

Generally, we compare the similarity of the O_{jl} with all the models trained for speaker i. For pass-phrase based verification, we only choose the (Θ_{il}) with the contents of the speech corresponding to the pass-phrase l. We get the score $\mathrm{S}_{jl\sim il}$ in place of the series of scores $\mathrm{S}_{jk\sim il}$ where $k = 1, 2, ..., K, l \in \{1, 2, ..., K\}$. The reason can be explained through an experimental observation illustrated in Fig. 5. There are four score distributions [4] in Fig. 5, these are the distributions with (a) genuine speaker speaking correct pass-phrase, (b) imposter speaker speaking correct pass-phrase, (c) genuine speaker speaking wrong pass-phrase and (d) imposter speaker speaking wrong pass-phrase. It can be seen that the log-likelihood ratio score with correct pass-phrase is most likely greater than that with wrong pass-phrase. It is also observed that the imposter trial with correct pass phrase (IC) is the closest distribution to the genuine trial with correct pass phrase (GC). It implies that the pass phrase is more informatively important than who the speaker is.

3. Pass-phrase based Speaker Verification System

An UBM can be denoted by the set of parameters, $\mathrm{u} = \{\bar{\omega}_i, \bar{\mathbf{m}}_i, \bar{\Sigma}_i; i = 1, 2, ..., C\}$, where C is the number of Gaussian components. The adapted GMM, λ, takes a similar form $\lambda = \{\omega_i, \mathbf{m}_i, \Sigma_i; i = 1, 2, ..., C\}$ where $\mathbf{m}_i, \Sigma_i, \omega_i$ are respectively the mean vector, the covariance matrix, and the weight of the ith Gaussian component.

The pass-phrase speaker verification system is a fusion of three subsystems including GMM-UBM, JFA and i-vector. Fig. 6 shows the diagram of the system. In the verification system, the three subsystems share exactly the same UBM, using the same feature extraction algorithm and subsequently share the same sufficient statistics. The weighting is trained through the development database for score normalization and calibration of each subsystem, then the scores from three subsystems are fused with the weights obtained by using the development database.

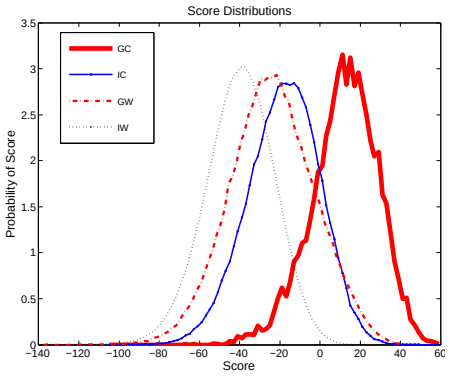

Figure 5: Score distributions of different situations. GC: Genuine trial with the same pass-phrase; IC: Imposter trial with the same pass-phrase; GW: Genuine trial with different pass-phrase; IW: Imposter trial with different pass-phrase.

3.1. GMM-UBM

In conventional MAP, λ is obtained by

$$\check{\lambda} = \arg\max_{\lambda} \left[f(\mathbf{X}|\lambda) g(\lambda) \right] \tag{1}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_\kappa]$ is the sequence of feature vectors, which we call the adaptation data. $\mathbf{x}$ is a J-dimensional feature vector. $f(\mathbf{X}|\lambda)$ is the likelihood of $\mathbf{X}$ given a GMM λ. $g(\lambda)$ is prior density of the GMM λ.

Assuming that the weights that are required to be a conjugate distribution are modeled as a Dirichlet density $g_1(\omega_1, ..., \omega_C)$ while mean and covariance of GMM is a conjugate prior distribution with normal-Wishart density $g_2(\mathbf{m}_i, \Sigma_i)$. $g(\lambda)$ is the joint prior density of g_1 and g_2. We have the mean and covariance parameters of the ith Gaussian adapted as follows [29],

$$\mathbf{m}_i = \alpha_i \check{\Xi}_i + (1 - \alpha_i) \bar{\mathbf{m}}_i \tag{2}$$

$\check{\Xi}_i$ is the first order sufficient statistics which has been normalized by the occupancy count; α_i are the adaptation coefficients given by

$$\alpha_i = \frac{N_i}{N_i + \gamma_i} \tag{3}$$

The relevance factor γ_i is a constant parameter in the normal-Wishart density as which the Gaussian parameters are modeled [29]; N_i is the occupation count which is directly proportional to the duration of the feature sequence.

The GMM-UBM system scores the test segment against the adapted GMM and the UBM models. The test score is given by the log likelihood ratio between the two models.

3.2. JFA

The JFA has been reported to have superior performance due to its robustness in channel compensation. In JFA, the speaker variability is modeled by the eigenvoice, where several common factors are used to represent the spanned space of the speaker, while the channel variability is modeled using a set of latent channel factors. In particular, a speaker-dependent

GMM-supervector can be decomposed in joint factors as follows [11]

$$\mathbf{m} = \bar{\mathbf{m}} + V\mathbf{v} + U\mathbf{u} + D\mathbf{d} \quad (4)$$

where $\bar{\mathbf{m}}$ is a speaker-independent supervector from UBM, V is the eigenvoice matrix, $\mathbf{v}$ is the eigenvoice factors (or speaker factors) with normal prior distribution; U is the eigenchannel matrix, and $\mathbf{u}$ the channel factors with normal prior distribution; D is the residual diagonal matrix, and $\mathbf{d}$ denotes the speaker-specific residual factors with normal prior distribution.

As a result of the decomposition in (4), speaker adaptation can be performed by updating a set of speaker-dependent latent variables and minimizing the influence of channel effects in an utterance. In our implementation, we train the eigenvoice matrix V by assuming U and D to be zeros; then train the eigenchannel matrix U given the estimate of V by assuming D to be zero; finally D matrix is trained given the estimates of V and U. In the training database design, for V matrix, we focused on obtaining the speaker-based principal dimensions; for the U matrix, the key is to obtain the channel (or nuisance) based principle dimensions. With the trained matrices V, U and D , the estimate of $\mathbf{v}$, $\mathbf{u}$ and $\mathbf{d}$ are obtained based on the posterior means given the particular utterance.

The score can be obtained by comparing the target speaker speech side and test segment statistics

$$S = (V\mathbf{v}_{tar} + D\mathbf{d}_{tar})^T \Sigma^{-1} (\Xi_{test} - N_{test}\bar{\mathbf{m}} - N_{test}U\mathbf{u}_{test}) \quad (5)$$

where $\mathbf{v}_{tar}$ and $\mathbf{d}_{tar}$ are the target speaker factors and residual factors; while Ξ_{test}, N_{test}, and $\mathbf{u}_{test}$ are the first order sufficient statistics, zero-order statistics (or occupation count), and the channel factors of the test speech utterance(s). We can see that the target speaker side is centered around speaker and residual factors, while the test speech has speaker-independent and channel factors removed. However, through experiments we notice that when the eigenvoice is not included but only consider eigen-channel factor and residual factors, the performance of the pass-phrase modeling is improved much. This is because the pass-phrase modeling actually separate the speaker into different model corresponding different phrases; this causes the eigenvoice is not constrained in the pur speaker modeling but in the speaker-phrase modeling, with very limit speech data, the eigenvoice can not be training properly in current database platform. Therefore, the score computation is re-written as follows

$$S = (D\mathbf{d}_{tar})^T \Sigma^{-1} (\Xi_{test} - N_{test}\bar{\mathbf{m}} - N_{test}U\mathbf{u}_{test}) \quad (6)$$

In score normalization, the z-norm and t-norm is used since they have been proven to effectively reduce the variability of the likelihood ratio scores that are used in the decision criterion.

3.3. i-Vector

Recently, Dehak et al. [14] proposed a feature extractor inspired by the JFA. Unlike JFA which models separately speaker and channel variability in a high dimension space of supervectors, the main idea is to find a low dimensional subspace projected from the GMM-supervector space, named the total variability space that represents both speaker and channel variability. The vector in the low-dimensional subspace is called i-vector.

The i-vector has been shown to respond well to generative modeling. Actually, the i-vector is estimated by evaluating the posterior expectation of the hidden variables in the model conditioned on the Baum-Welch statistics extracted from the utterance. This posterior calculation provides a posterior covariance matrix as well as a posterior expectation. The posterior covariance matrix can be interpreted as quantifying the reliability of the point estimate. An i-vector system uses a set of low-dimensional total variability factors w to represent each utterance. Each factor controls an eigen-dimension of the total variability matrix T. The total variability factors w is the i-vector. In particular, the GMM-supervector m can be decomposed into speaker-independent supervector $\bar{m}$ and the speaker-dependent supervector $T\mathbf{w}$

$$\mathbf{m} = \bar{\mathbf{m}} + T\mathbf{w} \quad (7)$$

To train T, just using the same procedure used for training V in JFA but treat all utterance of all training speakers as belonging to different speakers. Thus T actually absorbs the information of V, U and D in JFA. w is the latent variable. For each utterance, the i-vector ϕ is the posterior mean of w given an observation (or an utterance) $\mathbf{w}$ can be obtained given T.

In fact, i-vector extractors are trained without speaker-level labeling. It indicates that further transformations should apply in order to increase their speaker discriminative capacity. In i-vector system, a score can be obtained by comparing the enrollment i-vector and the test i-vector. It was shown that by projecting i-vectors onto a Linear Discriminative Analysis (LDA) basis, trained using representative enrollment data and speaker-labels to defined classes, the performance can be improved significantly. More effective performance can be obtained by giving the score with PLDA where the i-vector is considered as the second layer input vector to PLDA system [17].

There are two versions of PLDA named Gaussian and heavy-tailed versions. Currently, Gaussian PLDA [17, 18] and heavy-tailed PLDA [19], performed either on i-vectors directly or on the LDA-projected length-normalized i-vectors, yield state-of-the-art speaker recognition results. i-vectors can be approximately Gaussianized by length normalization so that the performance of Gaussian PLDA with length normalization is similar to that of heavy-tailed PLDA without length normalization. The recent research results show that unity length normalization of the i-vector indicates that Gaussian PLDA is as effective as heavy-tailed PLDA. In this investigation, we chose Gaussian PLDA for the speaker recognition.

In particular, given a speaker and a collection of i-vectors $\mathbf{w}_1 j, ..., \mathbf{w}_R j$ (one for each recording of the speaker in jth style (or channel or session)), standard Gaussian PLDA assumes that the i-vectors are distributed according to

$$\phi_{rj} = \varpi + \Omega\mathbf{h}_r + \Lambda\mathbf{q}_{rj} + \epsilon \quad (8)$$

incorporating speaker subspace Ω and channel subspace Λ. ϖ is the overall mean of the i-vectors. $\mathbf{h}$ and $\mathbf{q}$ are hidden variables representing the speaker factors and channel factors respectively; and they have standard normal priors. The residual ϵ_r is normally distributed with zero mean and diagonal covariance matrix. The PLDA is modeled by the parameters ϖ, Ω, Λ, and ϵ_r, which can be estimated through EM algorithm using the parameter training database. To make inference of the identity of a given test segment, the posterior probability for both enrollment i-vector and test i-vector generated from the same speaker or from different speaker are computed based on PLDA model. So, the log-likelihood ratio for the same and different inference likelihood is obtained as the output of the PLDA system. It has been proven that ignoring the channel subspace Λ and using full covariance matrix of ϵ_r instead of the diagonal matrix

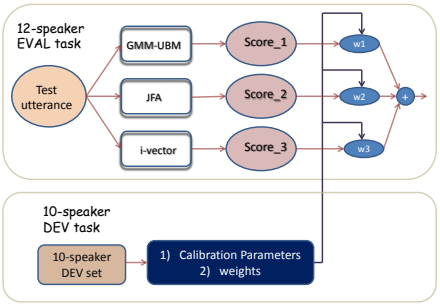

Figure 6: The Speaker Verification System.

Table 3: The composition of the Development (DEV) sets in terms of number of models (speaker-passphrase), number of train and test segments

Type of Phrase	#Model	#Train	#Test	#True	#False	Dur (second)
N	360	2120	1203	1203	13233	1.04
I	360	2130	1135	1135	12485	1.36
N+I	360	2120	1069	1069	11759	2.40

can be effective for speaker recognition system. Therefore the PLDA system in the investigation adopts this way. Finally, the S-norm is applied for score normalization [19].

4. Performance evaluation

In the evaluation, 19-dimensional MFCC coefficients, after voice activity detection (VAD), with their delta and double delta coefficients form the 57-dimension MFCC feature.

The dataset used for development (DEV) and evaluation (EVAL) consists of speech recordings from 22 speakers. The recording took place at two locations, i.e., in office of a port control center and on-board on the sea, where the communication between the two locations is through VHF wireless channel. Each speaker was required to read a list of names and IDs repeatedly. Given the dataset as described above, we split the 22 speakers into two sets consisting of 10 and 12 speakers respectively. The former is used to form the DEV set while the later is used to form the EVAL set. Table 3 shows the particular of DEV set while Table 4 gives the particular of EVAL set. The particulars listed in the tables include 1) **#Model**: total number of models trained, 2) **#Train**: number of training utterances used for model training, 3) **#Test**: number of utterances used for test, 4) **#True**: number of true trials for performance measurement, 5) **#False**: number of false trials for performance measurement, and 6) **Dur**: the average duration for each recording. All utterances for training were recorded in office environment, while all utterances for test were recorded on board. From the tables, the average duration of the 'NAME' is about 1.04 seconds, and the average duration for 'ID' is 1.36 seconds. Therefore, the combination of 'NAME' and 'ID' gives a total duration of 2.40 seconds in average.

The three subsystem (i.e., GMM-UBM, JFA and i-vector) share the same feature database and the same UBM with 256 mixture components. For JFA subsyetm, the joint factors are composed by 200 channel factors, and full rank diagonal ma-

Table 4: The composition of the Evaluation (EVAL) sets in terms of number of models (speaker-passphrase), number of train and test segments

Type of Phrase	#Model	#Train	#Test	#True	#False	Dur (second)
N	300	1800	850	850	7650	1.04
I	300	1800	890	890	8010	1.36
N+I	300	1800	790	790	7110	2.40

Table 5: Twelve speaker task: EER

Method	Name	ID	Name+ID
GMM-UBM	17.12%	17.68%	13.39%
JFA	15.73%	17.12%	11.69%
iVector	19.29%	20.79%	15.15%
fusion	14.63%	16.27%	11.13%

Table 6: Twelve speaker task: minimum DCF

Method	Name	ID	Name+ID
GMM-UBM	0.9310	0.8806	0.7619
JFA	0.9431	0.9392	0.8606
iVector	0.9194	0.8855	0.8657
fusion	0.9002	0.8511	0.7292

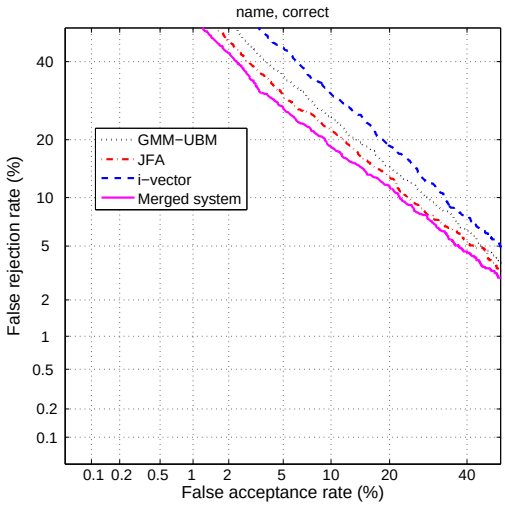

Figure 7: DET plot for VHF Name 12-speaker task.

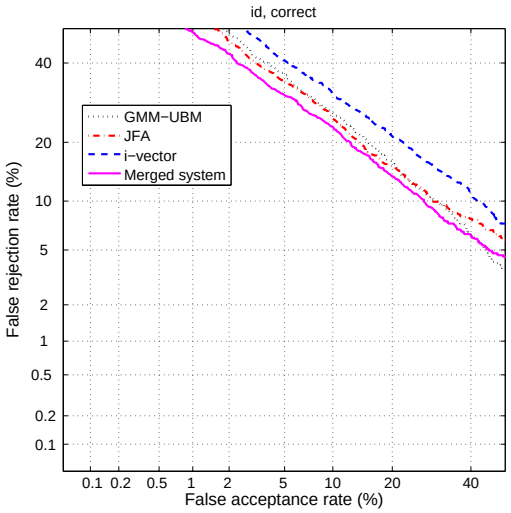

Figure 8: DET plot for VHF ID 12-speaker task.

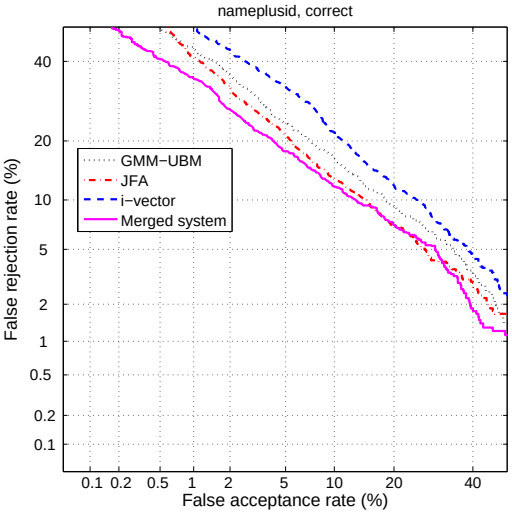

Figure 9: DET plot for VHF Name+ID 12-speaker task.

trix. For i-vector subsystem, the total variability is trained with 10 iterations. For the i-vector extractor matrix, 400 total variability factors are used; for PLDA training, 200 speaker factors are used. A fusion system is formed by calibrating each of subsystem and weighting the subsystems according to the trained parameters. The performance is evaluated on EVAL set while the DEV set is used to train various parameters including subsystem calibration and fusion weighting parameters. The performance is measured in terms of equal error rate (EER) and minimum detection cost function (minDCF) for the case of using 'NAME', 'ID' and 'NAME+ID' as pass-phrases.

Using 'NAME' or 'ID' only, the error rates are around 16% and 20%, respectively, on the DEV set. In order to improve the performance, both NAME and ID speeches are concatenated together so that the voice biometric system could make a better decision based on more speech information. Figs 7 and 8 show the DET curves of the subsystems and the fusion system for the 12-speaker EVAL task for the 'NAME' and 'ID' situations respectively. Fig 9 shows the DET curves with the subsystems and the fusion system result for the combination of 'NAME' and 'ID' for the 12-speaker EVAL task. Tables 5 and 6 show the EER and minimum DCF of the 'NAME', 'ID' and 'NAMD+ID' situation for the 12-speaker EVAL task. It can be seen from Table 5 that the error rate of 'NAME+ID' improves by 23% or 31.5% compared to the case where 'NAME' or 'ID' was used. We also can see that the expert system makes improvement at least 4.79% over any subsystem on the EVAL task.

We used the above mentioned JFA and i-vector algorithms on the NIST SRE 2012 platform, both sharing the same UBM and feature databases. Their performances in NIST SRE 2012 evaluation are effective in terms of EER, minimum DCF and actual DCF.

5. Conclusion

In this paper, we introduced a text-dependent speaker verification system, where the very short VHF speech was used. Against the short duration condition, a pass-phrase modeling concept was proposed. We analyzed the characteristics of the VHF and developed a fusion system consisting of GMM-UBM, JFA, i-vector for VHF speaker verification. Among the three subsystems, the UBM and the sufficient statistics were shared. According to the different conditions between enrollment and verification, we collected various databases and designed the suitable lists for various parameter training and final system setup. Especially, for the pass-phrase modeling, we noticed that JFA without channel factor consideration gives improved performance.

We investigated their performances in speaker recognition task in terms of EER and minimum DCF. The result shows that the fusion system gives advantage over any single subsystem. In the VHF platform, it has been observed that the state-of-the-art techniques: GMM-UBM, JFA and i-vector shows their mutual compensability. It is clear that, using longer speech, i.e. Name+ID, we could well achieve an error rate of less than 12%.

6. Acknowledgement

The works as reported in this paper were funded by MPA Maritime Innovation & Technology (MINT), Singapore.

7. References

[1] W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector Kernel and NAP variability compensation," *IEEE Intern. Conf. on Acoust., Speech, and Sig. Proce.*, ICASSP, vol. 1, pp. 97-100, Toulouse, 2006.

[2] C.H. You, K.A. Lee and H. Li, "An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition," *IEEE Signal Processing Letters*, vol. 16, no. 1, pp. 49-52, Jan. 2009.

[3] H. Li, B. Ma, K.A. Lee, H. Sun, D. Zhu, K.C. Sim, C.H. You, R. Tong, I. Karkainen, C-L Huang, V. Pervouchine, W. Guo, Y. Li, L. Dai, M. Nosratighods, T. Thiruvaran, J. Epps, E. Ambikairajah, E-S Chng, T. Schultz and Q. Jin, "The I4U system in NIST 2008 speaker recognition evaluation," *IEEE Intern. Conf. on Acoust., Speech, and Sig. Proce.*, ICASSP, pp. 4201-4204, Taipei, Apr. 2009.

[4] A. Larcher, K.A. Lee, B. Ma, and H. Li, "Phonetically-constrained PLDA modeling for text-dependent speaker verification with multiple short utterances," in Proc. TCASSP, 2013, pp. 7673-7677.

[5] D.A. Reynold and R.C. Rose, "Robust text independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, pp. 72-83, January 1995.

[6] B. L. Pellom and J. H. L. Hansen, "An efficient scoring algorithm for Gaussian mixture model based speaker identification," *IEEE Signal Process. Lett.*, vol. 5, no. 11, pp. 281-284, Nov. 1998.

[7] N. Brummer, L. Burget, J.H. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D.A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 15, pp. 2072-2084, Sep. 2007.

[8] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, pp. 308-311, 2006.

[9] C.H. You, K.A. Lee and H. Li, "GMM-SVM Kernel with a Bhattacharyya-based distance for speaker recognition," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 18, no. 6, pp. 1300-1312, Aug. 2010.

[10] K.A. Lee, C.H. You, H. Li, T. Kinnunen, and K.C. Sim, "Using discrete probabilities with Bhattacharyya measure for SVM-based speaker verification," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 19, pp. 861-870, May 2011.

[11] P. Kenny, "Joint factor analysis of speaker and session variability: theory and algorithms," CRIM, Montreal, *Technical Report, CRIM-06/08-13*, 2005.

[12] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, " A study of inter-speaker variability in speaker verification," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 16, no. 5, pp. 980-988, July 2008.

[13] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 15, no. 4, pp. 1435-1447, May 2007.

[14] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, Support vector machines versus fast scoring in the low dimensional total variability space for speaker verification, in Proceedings of Interspeech, Brighton, UK, Sep. 2009, pp. 1559-1562.

[15] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, " Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 19, pp. 788-798, May 2011.

[16] M. McLaren, and D. van Leeuwen, "Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors," *IEEE Intern. Conf. on Acoust., Speech, and Sig. Proce.*, ICASSP, pp. 5456-5459, Prague, May. 2011.

[17] S.J.D. Prince, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. International Conf. on Computer Vision*, pp. 1-8, Rio de Janeiro, Brazil, Oct. 2007.

[18] Y. Jiang, K.A. Lee, Z. Tang, B. Ma, A. Larcher, and H. Li, "PLDA modeling in i-vector and supervector space for speaker verification," in *Proc. Interspeech*, 2012.

[19] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Odyssey 2010 - The Speaker and Language Recognition Workshop*, 2010.

[20] P. Matejka, O. Glembek, F. Castaldo, M. Alam, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," *IEEE Intern. Conf. on Acoust., Speech, and Sig. Proce.*, ICASSP, pp. 48284831, 2011.

[21] http://www.mpa.gov.sg/

[22] National Institute of Standards and Technology, *NIST Speaker Recognition*, site available: http://www.itl.nist.gov/iad/mig/tests/spk.

[23] National Institute of Standards and Technology, "The NIST year 2012 speaker recognition evaluation plan," available: http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v11-r0.pdf.

[24] S. Furui, *Digital Speech Processing, Synthesis and Recognition*, 2nd Edition

[25] Boll S.F, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Tran. on Acoustics, Speech and Signal Processing*, ASSP-27, 2, pp. 113C120, 1979.

[26] H. Sun, B. Ma, and H. Li, "An efficient feature selection method for speaker recognition," *Proc. ISCSP*, pp. 181-184, 2008.

[27] A. Larcher, K.A. Lee, B. Ma, and H. Li, "The RSR2015: database for text-dependent speaker verification using multiple pass-phrase," *Proc. Interspeech*, 2012.

[28] A. Larcher, K.A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Comm.*, vol. 60, pp 56-77, May 2014.

[29] J.L. Gauvain and C-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 291-298, 1994.