PhD One-Year Progress Report

Speaker De-identification for Privacy Protection

Shufei He

Supervisor: Professor Ahmet Kondoz, Dr. Xiyu Shi

Institute for Digital Technologies

Loughborough University in London

September 2016

Table of Contents

Abstract

Within the framework of the speech processing technologies, speaker de-identification is a term used for concealing speakers' identities from their spoken utterances, while still preserving the intelligibility of the pure depersonalised linguistic information. Such techniques have a wide range of applications in real life speech-based scenarios when speaker anonymity protection is required including human-human conversations as well as human-machine voice interactions.

As a first step, we present a detailed comparison of the existing speaker de-identification solutions, followed by a list of challenges that need further research. The state-of-the-art techniques introduced in this report could be classified into two groups: i) methods based on voice conversion; ii) methods based on speech/diphone recognition and synthesis.

The first limitation observed from the literature review is the fact that the efforts of the researchers have been focused mainly on de-identifying spectral features of speech, but de-identifying prosodic characteristics of voice is still an important challenge. For this reason, in this report, a new speaker de-identification method based on linear predictive analysis is proposed, which combines both the spectral de-identification and prosodic de-identification, so that a better de-identification performance could be expected theoretically. Additionally, in contrast to other existing alternatives, the new method does not require previous training of the target speaker to be de-identified, thus it allows a more flexible online de-identification. In order to prove the validity of the new method, it is compared to the state-of-the-art systems based on statistical Gaussian mixture model, the most popular technique in the speaker de-identification world. The new speaker de-identification system and the baseline systems are evaluated by the state-of-the-art speaker recognition system. It is concluded that the new method outperforms baseline systems in terms of de-identification performance.

1. Introduction

1.1 Speaker de-identification: definition

''It's me!'' This claim is usually made over the telephone conversation or at an entrance out of view of the intended listener. It denotes the expectation that one's voice is sufficient for the listener to recognise the speaker. From a speech signal processing point of view, this can be explained as the fact that the speech signal carries massive amounts of speaker-dependent biometric information that allows the listener identifying the person who is speaking.

In this context, the central topic of this report, speaker de-identification, can be considered as a part of the speaker biometric information protection area. The goal of speaker de-identification systems is to conceal speakers' identities from their spoken utterances, while still preserving the intelligibility of the pure depersonalised linguistic information [1]. In other words, it is intended to enable privacy-preserved and security-assured speech transmission of ''what was said'' but to disguise traceable information of ''who said it''. Hence, the voice characteristics of the speaker have to be identified by the system and then modified or replaced by different voice features, without losing any information or modifying the message that is being transmitted. In general, an efficient speaker de-identification system has to be capable of accomplishing three main tasks:

- the system has to hide speakers' true identities from their spoken utterances;
- the system has to preserve voice quality of the de-identified speech;
- the system has to allow authorised listeners to retrieve the original identity of the source speaker and avoid un-authorised listeners to re-identify the source speaker.

1.2 Speaker de-identification: application

Speaker de-identification systems have a wide range of applications in real life speech-based scenarios when speaker anonymity protection is required. For example, speaker de-identification can be applied to protect the witnesses in investigations, whistle-blowers, and journalists' sources, etc. In addition, speaker de-identification can be utilised to preserve the privacy of individuals against their confidential medical information. Moreover, speaker de-identification can also be exploited to assure the freedom of expression and democracy in the public domains.

Cloud-based voice user interface, a new deployment of human-machine interaction, enables individuals convenient access to information exchange through spoken languages. Its benefits include not only hands-free and eyes-free, but also high communication efficiency and positive user experience. With all these charming benefits, cloud-based voice user interface is extensively used to activate, control, and operate numerous smart devices such as smart TVs, smart vehicles, smart air conditioners; and various smart personal assistants as well, such as Apple's Siri, Amazon's Echo, Microsoft's Cortana, Google's Now, and etc. In general, typical entities of a cloud-based voice user interface system include a speech-driven smart device, a cloud service provider, and third-party service providers, as shown in Figure 1.1. These smart devices and applications listen to all the sounds in users' immediate vicinity and send them back to the servers in the cloud for speech recognition and semantic analysis continuously. A wide variety of personal information might be contained in these voice recordings, for instance, user ID credentials, medical conditions, financial information, private conversations, and etc. Under this circumstance, users of cloud-based voice-driven services are increasingly concerned about the trustworthiness of such systems as well as the possible compromise of their privacy. For example, one of the fears is that users could be individually identified by the curious Cloud Service Providers (CSP), and their private conversation and sensitive information could be leaked out to the malicious Third-Party Service Providers (TPSP) and malicious outsiders. This is a core problem as individuals or enterprises concerned about their privacy are likely to refuse participation in such systems, and consequently to slow or stop the development of the cloud-based and speech-based smart living environments. In this context, it is desirable that speaker's voice identities could be removed before the speeches are transmitted to the cloud servers. Hence, speaker de-identification techniques could be applied to potentially improve the robustness of the cloud-based voice-driven user interface systems.

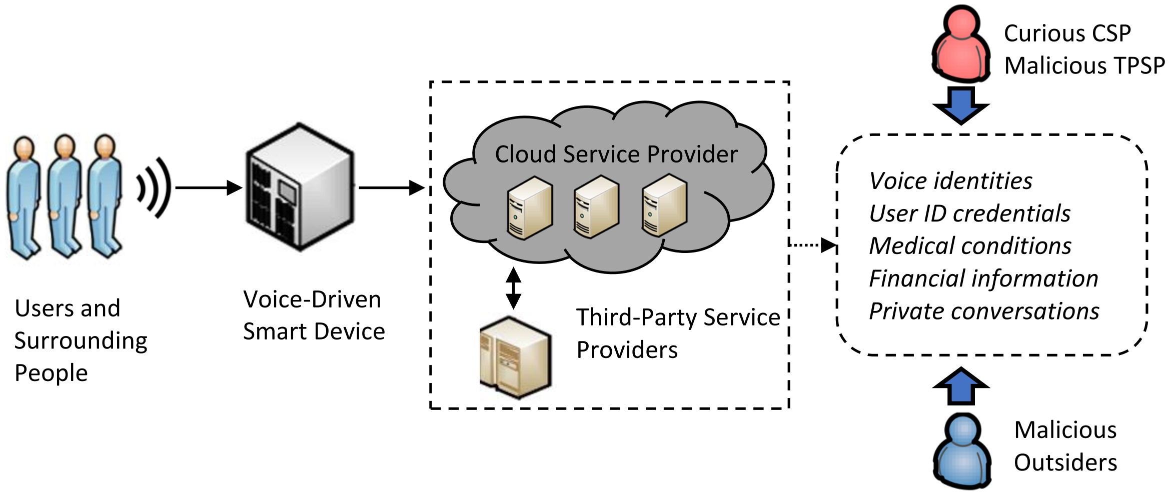

Figure 1.1 Framework of cloud based voice user interface system

Another application field of speaker de-identification is to enable the development of privacy-preserving technologies that use speech biometrics, such as the speaker verification systems. The objective of speaker verification is to accept or reject an identity claim from a speech sample [2]. Speaker verification systems have been applied to real-world security applications as access control solutions for more than one decade. Compared with the long, complex, and frequently changing passwords, speaker verification systems offer greater convenience to initiate the automated services through spoken languages, while maintaining sufficiently high accuracy and ensuring that the user are present at the point and time of verification [3]. The application of speaker verification techniques is increasing rapidly in a broad spectrum of industries nowadays, including the financial bank services, retail, corrections, even entertainment [4]. For example, a large amount of banks are deploying voice biometrics as their primary means to authenticate customers to their call centres, e.g. HSBC, Barclays Bank, Banco Santander Bank, Royal Bank of Canada, Tangerine Bank, Manulife Bank, and etc. [5].

However, speaker identity leakage poses serious threats to such systems, i.e. voice replay attacks and spoofing attacks with synthesised voice. For instance, voice conversion, which utilises voice biometrics to impersonate a particular speaker, can be intentionally used to deceive the voice-secured access systems [6]. With the help of voice conversion, one can gain unauthorised access to these voice-secured services with a synthetically generated voice. Multiple numbers of studies have been done to investigate the effect of transformed speech on speaker recognition systems [6][7][8][9]. These studies point out that correctly recognition rate decreases as the amount of conversion training data increases. Hence, voice conversion techniques could be deliberately used to fool the speaker recognition systems. In this context, users' privacy could be violated and their security could be broken. As a result, nowadays there is a major demand to develop the technology capable of concealing the speaker's identities in order to preserve users' privacy. This technology is referred to as speaker de-identification.

Finally, from a scientific point of view, acquiring a high level of knowledge about speaker individuality would be very useful to make progress in other speech processing technologies such as the speaker-independent speech recognition, speaker recognition, very-low-bandwidth speech coding using an adequate parameterization of the speaker-dependent information, etc.

1.3 Scope of work

The overall project consists of five interlinked work packages, with each one having specific outcomes to support others. Figure 1.2 illustrates the framework of the research diagram and the inter-connections between different works.

Work 1: Risk assessment against speaker identification

This work will identify security, privacy and trust issues against speaker identification in real-life voice-driven scenarios. Particularly, both the human-human conversations and human-machine voice interactions will be considered here.

Work 2: Literature review for speaker de-identification solutions

This work will survey into the existing speaker de-identification solutions. Based on the comparative analysis of these state-of-the-art techniques, research challenges that need future efforts in the speaker de-identification domain are to be concluded.

Work 3: High-quality speaker de-identification strategy

The general objective of this work is to design a speaker de-identification method with high quality and flexible versatility. Additionally, the state-of-the-art speaker de-identification systems will be implemented as baselines for further comparison with the novel high-quality strategy. The new acoustic de-identification method aims to accomplish five specific objectives:

- ❒ to enhance the efficiency of the speaker de-identification performance;
- ❒ to preserve the intelligibility of the de-identified speech signals;
- ❒ to assure any unseen speakers to be de-identified without the previous enrolment of their speech samples;
- ❒ to allow authorised listeners to retrieve the original identity of the speaker but to avoid un-authorised listeners re-identifying the speaker;
- ❒ to guarantee individuality between different speakers.

Work 4: Evaluation for speaker de-identification methodologies

This work will establish suitable evaluation porotypes for speaker de-identification methods and develop relevant computational tools for the performance experiments. Both the subjective experiment and objective experiment will be conducted here in order to assure the degree of reliability of the test results.

Work 5: Implementation of speaker de-identification system with voice UI

The objective of this work consists of the implementation of voice user interface (UI) system and the integration of the resulting speaker de-identification system into the implemented voice UI. It thus will lead to impact in real-world applications.

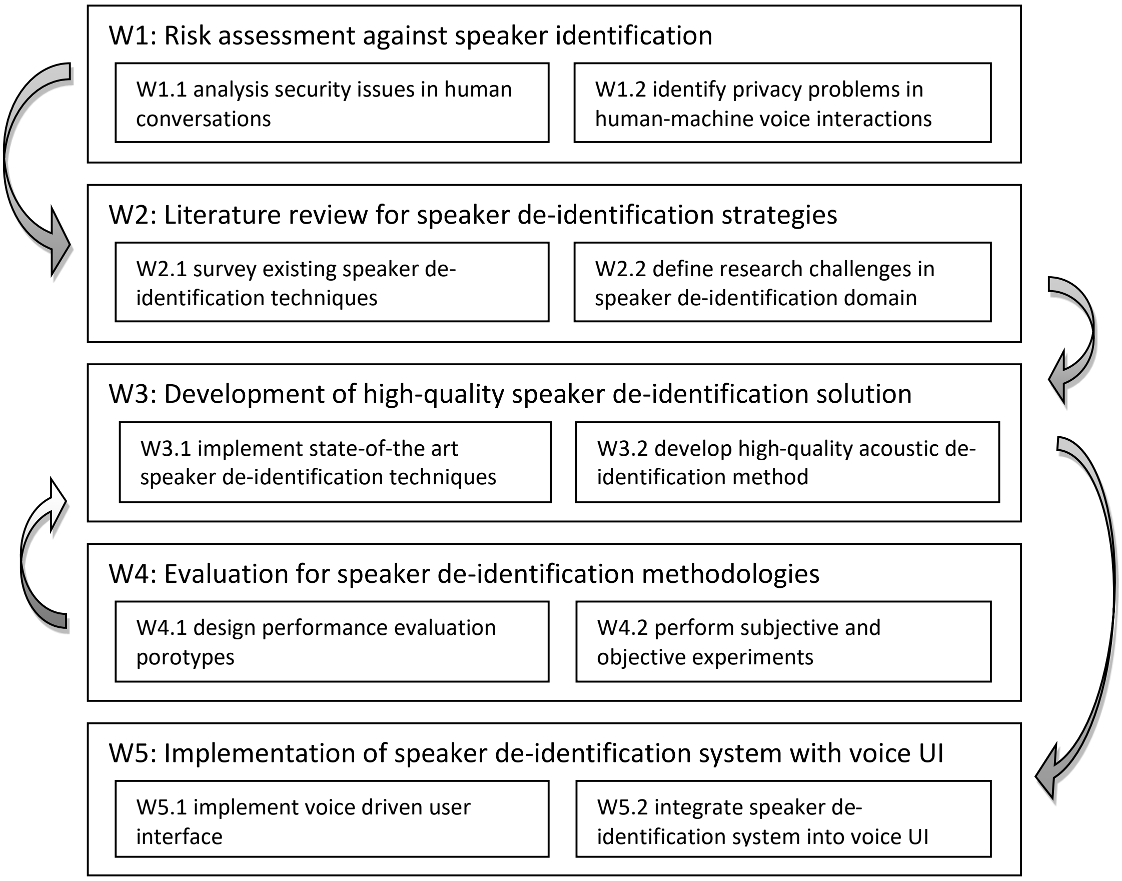
W1: Risk assessment against speaker identification
W1.1 analysis security issues in human conversations
W1.2 identify privacy problems in human-machine voice interactions
W2: Literature review for speaker de-identification strategies
W2.1 survey existing speaker de-identification techniques
W2.2 define research challenges in speaker de-identification domain
W3: Development of high-quality speaker de-identification solution
W3.1 implement state-of-the art speaker de-identification techniques
W3.2 develop high-quality acoustic de-identification method
W4: Evaluation for speaker de-identification methodologies
W4.1 design performance evaluation porotypes
W4.2 perform subjective and objective experiments
W5: Implementation of speaker de-identification system with voice UI
W5.1 implement voice driven user interface
W5.2 integrate speaker de-identification system into voice UI

Figure 1.2 Framework of research diagram

1.4 Report overview

The rest of the report is organised as follows.

In Chapter 2, various parameters and measurements used in the performance evaluation for speaker de-identification systems are outlined. The contribution of this chapter is a taxonomy of measurements in terms of five aspects:

- ❒ Efficiency of speaker de-identification
- ❒ Intelligibility of the de-identified speech
- ❒ Limitation on target speakers to be de-identified
- ❒ Reversibility and non-reversibility of de-identified speech
- ❒ Individuality between different speakers

Chapter 3 provides a detailed comparison of existing speaker de-identification solutions, followed by a list of research challenges that need future efforts. The state-of-the-art techniques introduced in this chapter could be mainly classified into two groups:

- ❒ Methods based on voice conversion
- ❒ Methods based on speech/phonetic recognition and synthesis

In Chapter 4, two baseline speaker de-identification systems are built using the harmonic plus stochastic model and the state-of-the-art voice conversion techniques. After that, a novel high-quality strategy combined with prosodic characteristics de-identification and spectral envelopes de-identification are proposed. It gives very good results in terms of de-identification performance, without the previous enrolment of speakers to be de-identified.

Chapter 5 is devoted to the research plans for future works, which can be divided into three main work packages:

- ❐ Development of high-quality speaker de-identification strategy
- ❐ Evaluation for speaker de-identification methodologies
- ❐ Integration of speaker de-identification system into the voice user interface system

Finally, in Chapter 6, the main conclusions of this report are summarized.

2. Evaluation criteria for speaker de-identification systems

This section presents a thematic taxonomy of performance evaluation measurements for speaker de-identification. These measurements can be basically categorised into five groups: i) efficiency of speaker de-identification, ii) intelligibility of the de-identified speech, iii) reversibility and non-reversibility of the de-identified speech, iv) limitation on target speakers to be de-identified, and v) individuality between different speakers.

2.1 Efficiency of de-identification

Efficiency assessment of speaker de-identification is the process of checking if the original speakers could be recognised from their de-identified speeches. Speaker identification system, speaker verification system, and also human recognition test are used to carry out the efficiency assessment for speaker de-identification [6][10][11]. Speaker identification is the method to determine the identity of the speaker who produced the input speech signal, while speaker verification is the method to accept or reject an identity claim from a speech sample. Generally, the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are used to measure the recognition accuracy of a speaker verification system. The higher FAR or higher FRR represents lower recognition accuracy as defined as:

$$FAR = \frac{\#false\ acceptances}{\#total\ verification\ attempts} \times 100 \tag{2.1}$$

$$FRR = \frac{\#false\ rejections}{\#total\ verification\ attempts} \times 100 \tag{2.2}$$

Here, the de-identification rate (DIR) is used to evaluate the efficiency of a speaker de-identification system, which is calculated as

$$DIR = \frac{\#total\ attempts - \#identified\ attempts}{\#total\ attempts} \times 100 \tag{2.3}$$

Particularly, the higher DIR stands for higher efficiency of the speaker de-identification system. For example, if the speakers can still be identified through their de-identified speech by speaker identification or speaker verification systems or human listeners, then we can conclude that the speaker de-identification solution does not perform as desired in speaker de-identification.

2.2 Intelligibility of de-identified speech

Intelligibility evaluation of de-identified speech is the examination to check if the non-linguistic speech content can be recognised and understood after linguistic de-identification. Speech recognition (SR) system and also human recognition test are used to evaluate the intelligibility of the de-identified speech. Here, the Word Error Rate (WER) is used to evaluate the intelligibility of a speaker de-identification system, which is calculated as

$$WER = \frac{\#unrecognised\ words}{\#total\ words} \times 100 \qquad (2.4)$$

Particularly, the higher WER represents lower intelligibility of the speaker de-identification system. For instance, if the context of de-identified speech can still be correctly recognised by speech recognition systems or human listeners, then it can be said that the speaker de-identification solution still preserves the intelligibility of the output de-identified speech.

2.3 Limitation on target speakers to be de-identified

Another performance evaluation aspect is the limitation on target speakers to be de-identified. In other words, it is the methodology to check whether the speaker de-identification system can de-identify all the new, unseen speakers. For example, if the speaker de-identification solution is only applicable to the speakers whose speech samples are acquired in advance for system training, then we can conclude that the speaker de-identification solution has considerable limitation on target speakers who can be de-identified.

2.4 Reversibility and non-reversibility of de-identified speech

Reversibility assessment of speaker de-identification is the experiment to check if the de-identified speech can be re-identified so that authorised listeners are able to retrieve the original identity of the speaker. In turn, non-reversibility assessment of speaker de-identification is the test to check if the de-identified speech cannot be re-identified so that un-authorised listeners are not able to access to the true identity of the speaker. To some extent, reversibility and non-reversibility are two opposite issues. On the one hand, a speaker de-identification system requires non-reversibility to assure secured speaker de-identification. On the other hand, it is also needed for the system to re-identify the source speaker in case that the real identity is required.

2.5 Individuality between different speakers

The individuality between different speakers is also an important characteristic of an efficient speaker de-identification system. Individuality assessment is the process of distinguishing different speakers after the phase of de-identification. Both speaker recognition systems and human listeners can be utilised to performance such experiments.

3. State of the art of speaker de-identification methods

Research studies on the speaker de-identification solutions have a relatively short history since the first speaker de-identification system was firstly proposed by Jin in 2009 [10]. A certain amount of techniques has been developed for speaker de-identification during the last decade. Despite the diversity of methods, they can be classified in two groups, depending on the type of de-identification that they apply: i) methods based on voice conversion; ii) methods based on speech/phonetic recognition and synthesis. A detailed explanation of the existing spectral conversion methods and algorithms is presented in the following subsections.

3.1 Methods based on voice conversion

3.1.1 Voice conversion framework

A vast majority of speaker de-identification solutions found in the literature apply voice conversion techniques to achieve the goal of speaker de-identification [7][10][11][12]. Voice conversion is a process of modifying the voice produced by a specific speaker, called source speaker, for it to be perceived by listeners as if it had been uttered by a different specific speaker, called target speaker [6]. The general framework of the voice conversion system is shown in Figure 3.1. The entire process can be divided into two modules: the offline training module and the runtime conversion module. In both modules, speech signals are firstly parameterised into short-term feature vectors, a process that is known as feature extraction. During the offline training phase, each source speaker's feature is paired up with phonetically equivalent target speaker's feature at frame-level, which is called frame alignment. Then, a conversion function is learnt from these source-target feature pairs, which map the utterance characteristics of source speaker to target speaker. In the runtime conversion phase, the learnt conversion function is used for converting source features to target features. Finally, the converted feature sequences are passed to a synthesis filter to reconstruct an audible speech signal. Consequently, the reconstructed speech signal contains different voice identities compared with the source speaker, so that we can say that the source speaker has been de-identified.

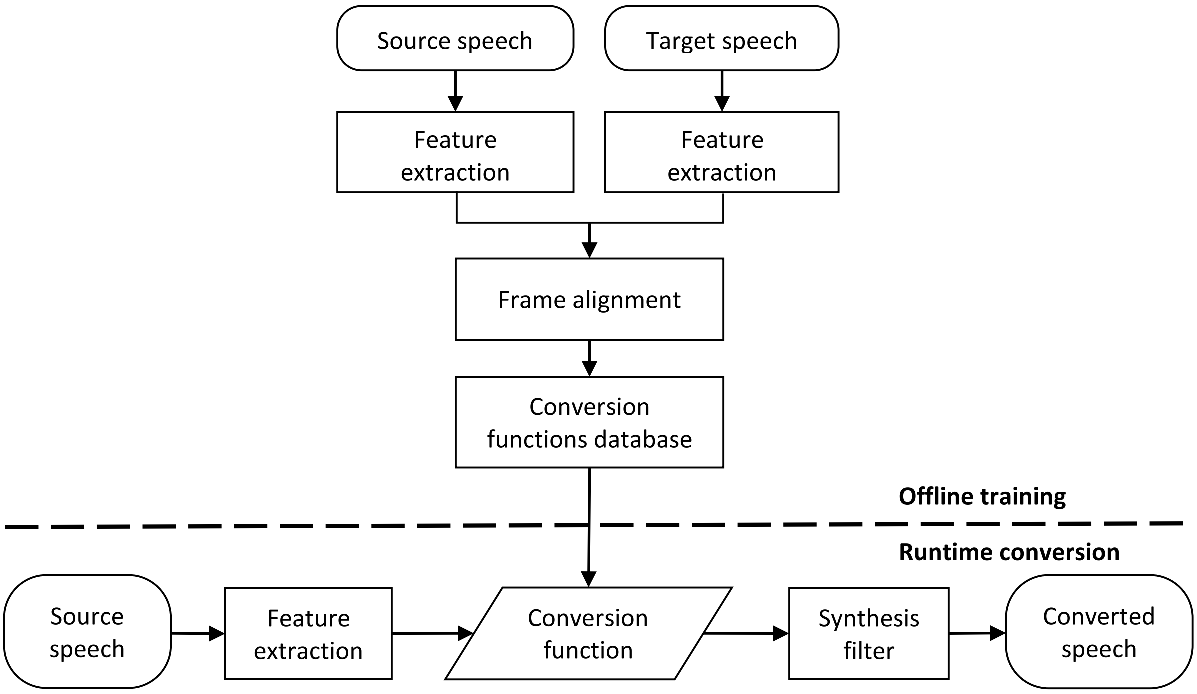

Figure 3.1 General framework of voice conversion system

3.1.2 GMM-based standard voice conversion

The earliest attempt to speaker de-identification starts from the idea whether a voice conversion system is able to deceive the speaker identification systems by Jin in 2008 [9]. Firstly, a set of conversion functions are trained between a closed set of source speakers and the target speaker with synthetic voice. Then, when a speech sample from the same set of source speakers is presented to the system, one of the conversion functions trained from the same speaker will be selected. Finally, the system can utilise the selected conversion function to transform the source voice to the target synthetic voice. The initial experiments showed that the GMM-based voice conversion system is able to fool the GMM-based speaker identification system with a relatively high de-identification rate of 92%, but not able to deceive the Phonetic-based speaker identification system with the de-identification rate of 42%.

In the GMM-based voice conversion systems, the Gaussian Mixture Model (GMM) based voice conversion technique was utilised to obtain the objective of speaker de-identification in Jin's work [9]. From the existing literature [11][12][11], [12], this GMM-based voice conversion approach later became the most widely used solution to speaker de-identification. Particularly, the acoustic mapping between source speaker and the target speaker is characterised by $\{x_t, y_t\}$, where the x_t denotes the source feature vectors and the y_t denotes the target feature vectors at frame t respectively. The joint Gaussian Mixture Model is fitted to these training acoustic vectors z_t as:

$$p(z_t) = \sum_{n=1}^{N} w_n^{(z)} N(z_t; \mu_n^{(z)}, \Sigma_n^{(z)}) \tag{3.1}$$

where $z_t = [x_t^T, y_t^T]^T$, T denotes the transposition of the vector, n and N demote the index and the total number of Gaussian mixture components respectively, $w_n^{(z)}$ denotes the weight assigned to the nth Gaussian component of z and $N(z_t; \mu_n^{(z)}, \Sigma_n^{(z)})$ denotes a Gaussian vector distribution defined by the mean vector $\mu_n^{(z)}$, and the covariance matrice $\Sigma_n^{(z)}$ of the nth Gaussian mixture component. These described parameters can be collectively represented by:

$$\lambda^{(z)} = \left\{ w_n^{(z)}, \mu_n^{(z)}, \Sigma_n^{(z)} \right\} \quad \mathrm{n} = 1, 2, \dots, N \tag{3.2}$$

$\mu_n^{(z)}$ and $\Sigma_n^{(z)}$ can be written as:

$$\mu_n^{(z)} = \begin{bmatrix} \mu_n^{(x)} \\ \mu_n^{(y)} \end{bmatrix} \quad \text{and} \quad \Sigma_n^{(z)} = \begin{bmatrix} \Sigma_n^{(xx)} & \Sigma_n^{(xy)} \\ \Sigma_n^{(yx)} & \Sigma_n^{(yy)} \end{bmatrix} \tag{3.3}$$

where $\mu_n^{(x)}$ and $\mu_n^{(x)}$ are the mean vectors of the nth component for source and target speaker respectively. The matrices $\Sigma_n^{(xx)}$ and $\Sigma_n^{(yy)}$ are the covariance matrices and the matrices $\Sigma_n^{(xy)}$ and $\Sigma_n^{(yx)}$ are the cross-covariance matrices of the nth component for source and target speaker respectively.

During the training phase, the Expectation-Maximisation (EM) algorithm is applied to estimate these model parameters which maximise the likelihood of the GMM distribution. It is worth noting that the GMM model parameters only represent the harmonic components of the speech frames here. For example, given a set of T training vectors $Z = \{z_1, \dots, z_T\}$, its GMM likelihood can be represented as:

$$P(Z|\lambda) = \prod_{t=1}^{T} p(z_t|\lambda) \tag{3.4}$$

During the testing phase, given an input vector x_t, the Minimum Mean Squared Error method [13] is used to predict the target vector y_t as:

$$F(x_t) = E[y_t|x_t] = \sum_{n=1}^{N} p_n(x_t) \left[\mu_n^{(y)} + \Sigma_n^{(yx)} \left(\Sigma_n^{(xx)} \right)^{-1} (x_t - \mu_n^{(x)}) \right] \tag{3.5}$$

where x_t is the LSF vector of the harmonic component at tth frame, $p_n(x_t)$ is the posterior probability that a given vector x_t belongs to the nth mixture component. The posterior probability $p_n(x_t)$ for component n is given by:

$$p_n(x_t) = \frac{w_n N(z_t; \mu_n^{(x)}, \Sigma_n^{(xx)})}{\sum_{m=1}^{N} w_m N(z_t; \mu_m^{(x)}, \Sigma_m^{(xx)})} \tag{3.6}$$

Finally, the stochastic component is predicted from the target speaker training vectors by means of linear transformation [14]. After the determination of conversion function for harmonic components, all the harmonic-stochastic vectors $\{y_t, y_t^s\}$ and the acoustic models of the target speaker given by $\left\{ w_n; \mu_n^{(y)}, \Sigma_n^{(yy)} \right\}$ can be used to calculate the optimal vectors $\{v_n\}$ and matrices $\{\Gamma_n\}$ of linear transformation that minimise the error of the following function:

$$y_t^s = \sum_{n=1}^{N} p_n(y_t) \left[v_n + \Gamma_n \left(\Sigma_n^{(yy)} \right)^{-1} (y_t - \mu_n^{(y)}) \right] \tag{3.7}$$

In conclusion, the mixture of Gaussian components is firstly used to model the probability densities of joint source and target speaker's feature vectors. Then a continuous probabilistic frame-wise mapping is applied to the source feature vectors. It can capture the overall spectral characteristics and can produce the average representation of the target spectrum. As a result, GMM based voice conversion can generate converted speech with good similarity to target speech, but with degraded quality [6].

3.1.3 GMM-based voice conversion combined with de-duration method

In later research, Jin et al. proposed to apply standard voice conversion combined with de-duration method to improve the de-identification performance[7]. The concept of duration of a sound is defined as an amount of time interval that a sound lasts. Thus, in this work, the goal of speech de-duration is obtained by applying consistent duration statistics to utterances regardless of speakers. The reason by performing the de-duration method is to eliminate the possibility that the duration statistics of the speaker might be exploited by the speaker recognition system to recognise the speaker.

The de-identification performance of the modified speech by using de-duration method is reported to be better than of the converted speech by using standard GMM method. In particular, it is reported that the de-identification rate of the modified speech by de-duration method is 96% against the GMM-based speaker identification system, and 46% against the Phonetic-based speaker identification system, while the de-identification rate of the modified speech by using standard GMM method is 92% against the GMM-based speaker identification system, and 42% against the Phonetic-based speaker identification system.

3.1.4 GMM-based voice conversion with chained transformation method

Considering the better performance of de-identification with the de-duration method compared to the standard GMM-based method described in the previous paragraph, a very simple approach chained with both the de-duration method and the GMM-based standard method was also presented in [7]. In this work, the output speech of the de-duration system is used as the input speech for the baseline standard voice conversion system.

From the experimental results in [7], the transformed speech with chained conversion method outperforms the de-duration method, with a de-identification rate of 67% against the Phonetic-based speaker identification system, while achieves same de-identification rate of 96% against the GMM-based speaker identification system.

3.1.5 GMM-based voice conversion combined with transterpolated concept

The appearance of previous GMM-based systems didn't lead to satisfactory results against the Phonetic-based speaker identification system, so Jin in [7] also proposed another approach based on the concept of transterpolation. By ''transterpolation'', it refers to the process of interpolation or extrapolation between source feature and target feature. In particular, the transterpolated feature, x, is computed as $x = s + f(v - s)$, where s is the value of the source speaker's feature, v is the value of the converted feature, and f is the factor of inter- or extrapolation. The intelligibility of the converted speech is reported to be very high with a speech recognition rate of 100% while the inter- or extrapolation factor f is in the range of 1.2 to 1.6.

It is reported in [7] that the transterpolated method outperforms previous three systems (the standard voice conversion method, the de-duration method and the chained conversion method). In particular, the transterpolated method with inter- or extrapolation factor f of 1.6 gives the best speaker de-identification performance, achieving de-identification rate of 100% against the GMM-based speaker identification system and 87.5% against the Phonetic-based one. Hence, it can be concluded that the GMM-based voice conversion system combined with transterpolated method could fool the conventional GMM-based speaker identification system, but performs undesired result against the state-of-the-art Phonetic-based speaker identification system.

3.1.6 GMM-based voice conversion combined with speaker model adaption

In [11], a speaker model adaption based approach was used to achieve the online speaker de-identification system in 2014. The definition of ''online'' is that any new and unseen speakers can be de-identified in the runtime without their previous enrolment when using the de-identification system. The framework of this work is shown in Figure 3.2. Firstly, a set of voice conversion functions are pre-learnt from a closed set of source and target speakers. When a speech sample from a new speaker is presented to the system, the system will perform a log-likelihood ratio examination between input source speaker and trained closed set of speakers through a speaker identification system. One voice conversion function would be selected which achieves the maximum likelihood with the input speaker. Finally, the system will apply this conversion function learnt from that trained speaker to the speech sample of the new speaker. In other words, an already trained conversion function between source speaker A and target speaker B is adapted to the acoustic data of a different source speaker C here. It is worth noting that the conversion functions used in this work are also based on Gaussian Mixture Model. The de-identified speech in this method is reported to achieve the comparable de-identification performance compared to the GMM-based standard voice conversion system against the GMM-based speaker identification system, but with added flexibility which ensures non-limitation among the target speakers to be de-identified. Particularly, the de-identification rate achieved by the speaker model adaption method is 87.4%, while it is 91% by the GMM-based standard voice conversion method [10]. Nevertheless, it is unclear from [11] that whether the speaker model adaption system can deceive the state-of-the-art speaker identification systems, such as the Phonetic-based one.

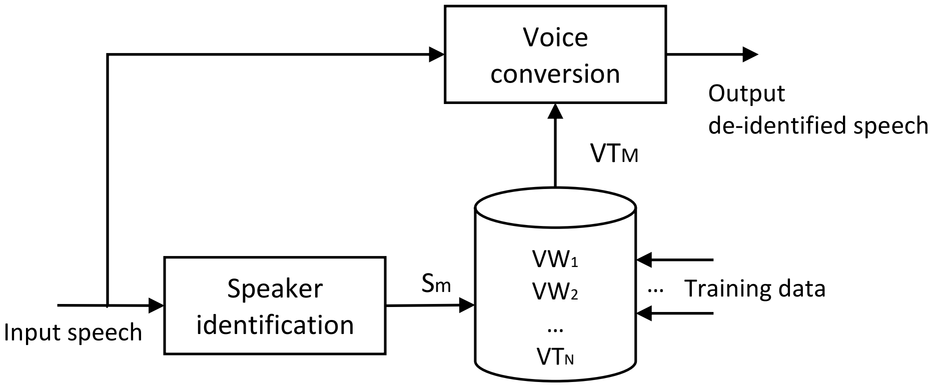

Figure 3.2 Framework of the model adaption based approach

3.1.7 GMM-based voice conversion combined with speaker selection

Recently in 2015, a discriminative approach for target human speaker selection applied in speaker de-identification system is presented in [12]. In this work, the most appropriate target speaker is selected by using a speaker identification system to achieve three specific goals: i) gives the lowest identification confidence to be identified as the source speaker; ii) does not converge to a certain speaker completely, but gives as much doubt as possible about the speaker identity, and iii) can achieve a desired result if the de-identification operation is reversed for the purpose of re-identification. Figure 3.3 illustrates the general framework of the speaker selection process. The whole calculation process can be expressed as the following formula:

$$K_i = \arg\max_{k\in K}\{-\alpha f(i,k) + \beta c(i,k) + \gamma d(i,k)\} \tag{3.8}$$

where K is the total number of speakers in the repository, i and k are the speaker index of source speaker and target speaker respectively, $f(i,k)$ is the identification confidence of the transformed speech to the source speaker i, $c(i,k)$ is the confusion factor of the transformed speech from the source speaker i to the target speaker k, $d(i,k)$ is the identification confidence of the re-transformed speech to the source speaker i, and α, β and γ are the weights of the previously described functions respectively. Additionally, two different confusion factors are defined for speaker selection: entropy and Gini index. The confusion factor for a source speaker i transformed to be a target speaker k using entropy or Gini index measure is calculated as: $c(i,k) = -\sum_{j=1}^{N} p_j\, log(p_j)$, $c(i,k) = 1 - \sum_{j=1}^{N} p_j^2$, respectively. Here, N is the number of speakers in the repository and p_j is the identification confidence that the transformed speech from i to k is recognised as the voice of speaker j. The speaker-selection based method is reported to have significantly better performance among the de-identification and re-identification performance than the method using synthetic voice.

Particularly, the speaker-selection based method achieves an average de-identification rate of 90.19% against the conventional GMM-based speaker identification system and 86.22% against the state-of-the-art i-vector-based one, while the baseline standard voice conversion method gives an average de-identification rate of 61.83% and 64.73% respectively as reported in [12].

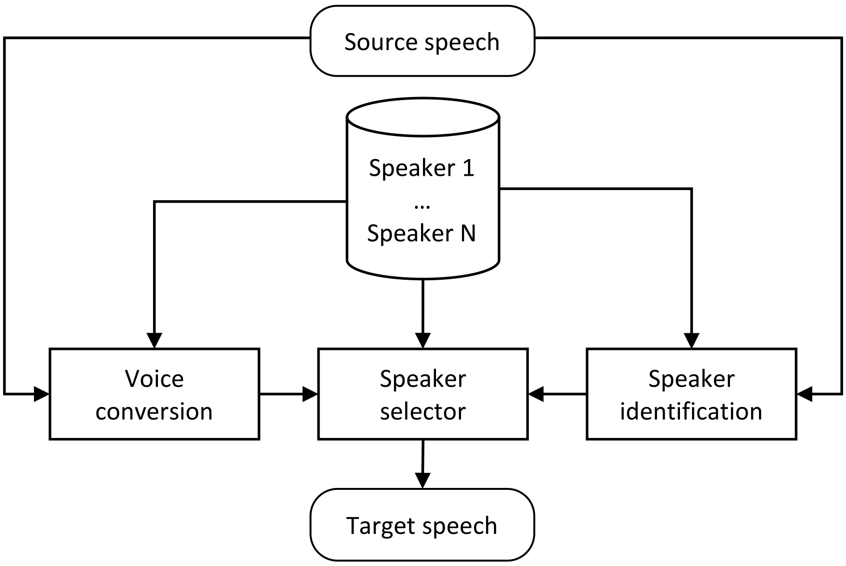

Figure 3.3 Framework of the speaker selection approach

3.2 Methods based on speech/phonetic recognition and synthesis

3.2.1 Phonetic recognition and synthesis system

In a very recent work in 2015, a novel approach based on phonetic recognition and synthesis has been proposed to optimise the speaker de-identification performance by Justin [1]. The block diagram of this method is shown in Figure 3.4. The system firstly recognises the input speech samples with a diphone recognition system, which is referred as the Hidden Markov Model Toolkit [15] and then synthesis the obtained phonetic transcription into the speech samples.

Considering different target speaker characteristics using different speech synthesis techniques, two speech synthesis systems are applied for the speech reconstruction process in this work to compare speaker de-identification performance and quality of the re-synthesised speech, including the HMM-based synthesis [15] and the diphone-based synthesis [16]. The HMM-based speech synthesis technique is based on Hidden Markov Model (HMM) [15], where the frequency spectrum, fundamental frequency, and duration of speech are modelled simultaneously by HMMs. In this system, speech waveforms are generated from HMMs themselves based on the maximum likelihood criterion. On the other hand, the diphone-based speech synthesis technique is based on the concatenation of diphones, which produces more natural-sounding synthesised speech but contains sonic glitches in the output speech due to the concatenative synthesis.

The speaker de-identification approach in [1] allows for the recognition and synthesis module to be trained separately without any previous enrolment and any speech material of the source speaker, and is capable of running in real-time. Due to the independence between the acoustical models of the recognition and synthesis modules, the phonetic recognition and synthesis approaches ensures the highest level of de-identification. The speaker de-identification test results in [1] against the speaker verification suggest that the recognition performance of both two synthesis implementations are more or less random. Thus, the phonetic recognition and synthesis based approach in this work is reported to be outperforming the previous works by using voice conversion techniques [7][10] in terms of speaker de-identification performance. However, the main disadvantage of the method is the intelligibility degradation of the re-synthesised speech, which is related to the performance of the speech recognition module. The recognition errors occurred in the recognition phrase lead to serious information losses of the output synthesised speech, which significantly decrease the intelligibility of the de-identified speech. Particularly, it is reported from the speech recognition experiments that average Word Error Rate (WER) is 33% for the HMM-based synthesis approach, while it is 21% for the diphone-based approach.

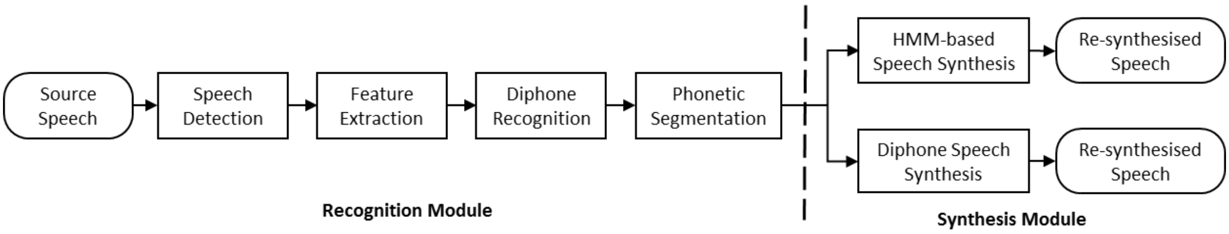

Figure 3.4 Block diagram of phonetic recognition and synthesis approach

3.2.2 Speech recognition and synthesis system

Another obvious solution for speaker de-identification is speech recognition and synthesis. Speech recognition is the methodology to convert a speech signal to a sequence of words, while speech synthesis is used to convert text into speech [17]. A number of studies have been done to develop speech recognition and speech synthesis techniques. One of the early concatenate speech recognition systems, the SPHINX system, was suggested by Lee et al. in 1990 [18]. An example of more practical applications using speech recognition techniques is autonomous broadcast news transcription system developed by Woodland et al. in 1999 [19]. Additionally, several well-defined convenient APIs are also available from websites. For example, Sun/Oracle’s Java Speech API (JSAPI) (1998) [20] and Microsoft’s Speech Application Software Developer's Kit (2005) [21] provide developers with APIs for speech recognition and speech synthesis. Moreover, notable speech processing toolkits including audio APIs are the Hidden Markov Model Toolkit (HTK) [15] developed for speech recognition systems and Festival Speech Synthesis System [15].

Theoretically, by using independent speech-to-text and text-to-speech techniques, the speaker de-identification approach can achieve reliable and efficient de-identification performance. Nonetheless, speech recognition system consumes significant memory and computational resources in local devices, which brings limitations in practical speaker de-identification applications. Moreover, the speaker de-identification also requires full-fledged and error-free speech recognition in order to preserve the speech content [7]. For example, the offline speech recognition system on mobile devices implemented by Google in 2016 [22] has a memory footprint of 20.3 MB, and achieves 13.5% word error rate on a dictation task. It can be concluded that the state-of-the-art offline speech recognition techniques nowadays do not perform as desired on mobile devices in terms of recognition rate. Therefore, speech recognition technique customized for local devices needs further efforts to implement the speaker de-identification solution.

3.3 Limitations and challenges

Based on the details of existing speaker de-identification solutions presented in this report, it is evident that during the last decade, the performance of speaker de-identification solutions has not reached a satisfactory level. This assertion is confirmed by the fact that the de-identification rates against the state-of-the-art speaker identification systems are not desired from the literature review. In order to address the research challenges in this promising area, here are some problems that still need further efforts:

1) The efforts of researchers have been focused mainly on de-identifying acoustic spectral features of speech; however, de-identifying prosodic characteristics of voice is still an important challenge. For example, most of the reviewed speaker de-identification solutions are based on the GMM voice conversion technique, which aims to only transform the magnitude spectrum of the frequency response of the vocal tract system. In fact, a better knowledge about prosody de-identification is essential for de-identifying non-neutral speeches. On the other hand, it is evident

that a complete speaker de-identification system should consider both spectral and prosodic features but, at present, this higher-level problem has not been addressed yet.

2) Currently, no scheme exists that achieves good scores on both de-identification performance and intelligibility of the de-identified speech. For example, phonetic/speech recognition and synthesis approaches are characterized by good de-identification scores but lower intelligibility scores, whereas systems using voice conversion methods do not significantly degrade the quality of speeches but are not good at de-identifying the voice. Further improvements are necessary to develop new methods that successfully conceal the identity of speakers but also minimise the speech quality degradation. This is important for real-life speaker de-identification applications in which voice-driven devices are expected to recognise the speech content and human listeners are expecting to hear natural-sounding voices without information losses.

3) The state-of-the-art schemes for speaker de-identification require system training and previous enrolment of speakers which brings limitations in real-world applications. In other words, by using the speaker de-identification system, users' speech materials need to be provided and their voice features need to be identified in advance. In most cases, such voice conversion based solutions require parallel corpora of aligned sentences for system training, with the same text spoken by both the source and target speakers. This also brings limitations in practical applications, where the user is unwilling to be identified by the system, e.g. in anonymous police systems or helplines. In addition, with only a closed set of speakers to be de-identified, it would be also not practical in applications with large numbers of potential users, e.g. call centres providing services to the general public. Hence, it is very desirable to develop online acoustic de-identification techniques that could be compatible with such applications, so that any unseen new users are able to use the speaker de-identification system, without providing speech samples for feature training in advance. Although few methods have been proposed recently to address the claimed issue [1], they have important disadvantages like the losses of speech content or their negative impacts on the de-identification scores.

4) No existing scheme is capable of assuring both reversibility and non-reversibility of the de-identified speech. In some practical applications such as telephone bank services, the de-identified speeches need to be re-identified by authorised listeners. In the meantime, these systems also need to assure that unauthorised listeners are not able to re-identify the speaker's identity. It is evident that how to design a scheme to achieve both reversibility and non-reversibility speaker de-identification is still quite challenging. For example, the voice conversion based approaches aim to learn a conversion function between source speaker and target speaker, which assures good reversibility of the de-identified speech. However, such methods also pose potential security issues if the conversion function is learnt by adversaries. Due to the independence between the recognition and the synthesis modules, the speech/diphone recognition and synthesis based solutions assure high level of non-reversibility of de-identification. Though, such systems are not applicable to the scenarios where the reverse process, to obtain the speaker's real identity, is required. So it is very desirable to develop a de-identification solution which allows authorised user to transform the de-identified speech back to the original speech but to a certain extent, avoiding unauthorised user to re-identify the speaker.

5) It is not possible to distinguish different speakers after de-identification at the present as all speeches are re-synthesised to the same target speaker. This brings limitations in which speaker individuality is required in some real-world applications, such as voice-driven smart assistants. In this case, it is desired if users can be individually recognised by voice-driven devices so that specific services can be provided for different users, e.g. personal recommendations.

6) The schemes need to be evaluated further using larger speaker databases in order to verify the accuracy of evaluation experiments. From the literature review, all the reviewed speaker de-identification systems were tested using speech corpus with a relatively limited number of speakers. For example, the voice conversion approaches in [7] was tested among the Wall Street Journal (WSJ0) corpus [23] which includes 24 male speakers. The speaker de-identification solution proposed in [11] is evaluated among a Croatian speech database of weather forecast [24],

which includes 10 male speakers. Hence, it is desirable that the de-identification solutions are evaluated using large speaker databases.

According to the objectives of the report, solutions for the claimed problem 1 and problem 3 are proposed in the following Chapter 4.

4. Initial works toward speaker de-identification

During the first year, several algorithms have been implemented, developed and tested with the aim of achieving online high-quality speaker de-identification. This chapter gives a description of the initial research work and is structured as follows.

- ❒ In Section 4.1, a state-of-the-art performance GMM-based speaker de-identification system is implemented as the baseline method for further experiment and comparison with the new method.
- ❒ Section 4.2 introduces and discusses the evaluation of a novel speaker de-identification technique, which improves the de-identification performance without the previous enrolment of the target speakers that are to be de-identified.
- ❒ Section 4.3 contains the objective speaker recognition experiment results and discussions on both the baseline system and the new method.

4.1 Baseline systems based on GMM and WFW

We use two different voice conversion techniques as the baseline systems: the standard Gaussian Mixture Model (GMM) based voice conversion system [25] and the Weighted Frequency Warping (WFW) based voice conversion system [26].

According to the bibliographic analysis carried out in Chapter 3, GMM based statistical conversion method [25] is the most reasonable choice as it is the only scheme applied for speaker de-identification approaches within the voice conversion domain. GMM based voice conversion generates converted speech with good similarity to target speech. Hence, we use the GMM based system as one baseline system for the purpose of de-identification performance comparison. However, GMM-based method does not perform desired voice quality due to the speech degradation in the output speech. Given that the WFW based voice conversion maintains good quality of converted speech, we also use the WFW based method as another baseline system.

Although the techniques implemented in this section do not contain relevant novelties with respect to the state of the art, the implementation of the GMM based and the WFW based voice conversion systems are needed as the baseline systems for further comparison with the newly proposed de-identification method in section 4.2. In the next sub-sections, the implementation details of the baseline systems are described according to the structure of a generic voice conversion system, shown in Figure 3.1.

4.1.1 Speech analysis and frame alignment

The first step before creating a voice conversion system consists of choosing a suitable speech model with certain properties. In this baseline systems, the Harmonic plus stochastic model (HSM) [27] is used to represent the speech signal frames. The HSM model assures flexible and high-quality modification among the prosodic features and spectral characteristics from the speech signals [28]. In this work, the HSM parameters include the fundamental frequency, the harmonic components and the stochastic components. During the speech signal analysis phase, the mean and variance statistics of pitch information is to be converted. Additionally, the amplitude and phase of the harmonic components below 5 kHz are translated to an all-pole filter. This filter is then converted in to its associated line spectral frequencies (LSFs). Finally, the stochastic components are represented by the LPC coefficients.

Additionally, the non-parallel alignment method [29] is employed in this baseline systems for the purpose of frame alignment. By non-parallel, we mean that different set of utterances from different speakers are used for the voice conversion training.

4.1.2 Pitch conversion

The fundamental frequency f_0 or pitch is one of the most important excitation features when considering the identity of one speaker. A logarithmic Gaussian normalized transformation is used for pitch level conversion [14].

$$logf_0' = \mu_{logf_0}^{y} + \frac{\sigma_{logf_0}^{y}}{\sigma_{logf_0}^{x}}(logf_0 - \mu_{logf_0}^{x}) \tag{4.1}$$

where $\mu_{logf_0}^{x}$ and $\sigma_{logf_0}^{x}$ are the mean and variance of f_0 in the log domain for the source speaker, $\mu_{logf_0}^{y}$ and $\sigma_{logf_0}^{y}$ are the mean and variance of f_0 in the log domain for the target speaker.

4.1.3 Spectral mapping using GMM

After non-parallel frame alignment, the spectral mapping between source speaker and the target speaker is characterised by $\{x_t, y_t\}$, where the x_t denotes the source feature vectors and the y_t denotes the target feature vectors at frame t respectively. The spectral mapping method by using GMM has been described extensively in Section 3.1, so no more details are given here.

4.1.4 Spectral mapping using WFW

Weighted frequency warping (WFW) is the combination of GMM-based and dynamic frequency-warping (DFW) based voice conversion methods [26]. During the training phase, the linear frequency warping function $W_n(f)$ of GMM component n is determined by the position of the formants. During the testing phase, the warping function $W_n(f)$ of GMM component n is determined as a linear combination of the N warping functions. It can be expressed as:

$$W^{(t)}(f) = \sum_{n=1}^{N} p_n(x_t)\, W_n(f) \tag{4.2}$$

where $p_n(x_t)$ denotes the posterior probability for GMM component n, given by Equation (4.6).

On the one hand, the input frame envelope $S^{(t)}(f)$ is warped in frequency domain as follows:

$$S_{dfw}^{(t)}(f) = S^{(t)}\left[W^{(t)}(f)\right]^{-1} \tag{4.3}$$

On the other hand, a converted spectrum $S_{gmm}^{(t)}(f)$ is produced by means of the GMM. The GMM transformed envelope captures overall trends in spectral power, while the DFW transformed envelope maintains spectral details. Hence, these two separate converted envelopes are then weighted in order to smoothen the spectral details of the frequency warped envelope.

Finally, an energy correction filter is applied that is a smoothed convolution of the ratio of GMM-transformed envelope and DFW-transformed envelope with a triangular in frequency window. Therefore, the final converted spectrum is provided here below:

$$S_{wfw}^{(t)}(f) = \left|\frac{S_{gmm}^{(t)}(f)}{S_{wfw}^{(t)}(f)}\right| * B(f)\, S_{dfw}^{(t)}(f) \tag{4.4}$$

Where * denotes the convolution operation and $B(f)$ is the triangular smoothing-in-frequency function of the energy correction filter. Moreover, the same prediction function given in Equation (3.6) is used to obtain the stochastic part of the converted envelope.

4.2 A new speaker de-identification method based on LPC

According to the objectives defined at the beginning of this report, this section presents a new acoustic de-identification method whose goals are i) to enhance the efficiency of the speaker de-identification performance; and ii) to allow unseen speakers to be de-identified without the previous enrolment of their speech samples.

As it was mentioned in Chapter 3, so far the efforts of researchers have been focused mainly on de-identifying spectral acoustic features of speech. However, prosodic characteristics of voice also contain important information about the speaker individuality. In this work, a new acoustic de-identification method based on spectral envelope de-identification combined with prosodic modification is proposed. Theoretically, this combination brings together the advantages of both approaches, so that a better de-identification performance could be expected.

4.2.1 Fundamentals of source-filter speech model

The human speaking process can be described by the so called source-filter model, where the source signal represents the airflow coming from the glottis, and the physical vocal tract is represented by a filter that modifies the frequency-shape of the source signal [14]. In this work, the source-filter speech model is selected to create the new speaker de-identification method. The reason is that theoretically, parameterising both the glottal source and the vocal tract allows the highest level of capturing the identifiable features from the speech.

From a signal processing point of view, in implementation of the source-filter model of speech production, the sound source or excitation signal $x(t)$ is often modelled as a periodic impulse train for voiced speech, or white noise for unvoiced speech. The vocal tract filter is approximated by an all-pole filter in the simplest case with impulse response $h(t)$, where the coefficients are obtained by performing linear prediction to minimize the mean-squared error in the speech signal to be reproduced. Convolution of the excitation signal $x(t)$ with the filter's impulse response $h(t)$ then produces the synthesised speech $y(t)$. The signal processing representation of the source-filter model is illustrated in figure 4.1.

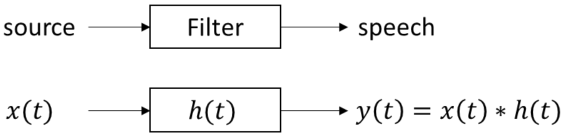

Figure 4.1 Signal processing representation of the source-filter model

Therefore, in the spectral domain, a speech signal Y of n frames $\{y_1, \dots, y_n\}$ can be represented as $Y(\omega) = H(\omega) \cdot X(\omega)$, where $H(\omega)$ represents the vocal tract transfer function of signal $Y(\omega)$, and $X(\omega)$ denotes the Fourier transform of the glottal source excitation signal [30]. Hence, a de-identified speech signal $Y'(\omega)$ can be obtained by replacing a de-identified transfer function $H'(\omega)$ and a de-identified source excitation signal $X'(\omega)$, as $Y'(\omega) = H'(\omega) \cdot X'(\omega)$. Particularly, it is worth noting that the transfer function is usually referred to as spectral characteristics, and the glottal source excitation signal is usually referred to as prosodic characteristics. A detailed explanation of both the prosodic de-identification and spectral de-identification is presented in the following subsections.

4.2.2 Pre-processing of speech signals

Pre-emphasis

The first stage in the signal re-processing is to boost the amount of energy in the high frequencies before the linear-prediction analysis, which is referred as pre-emphasis. During the reconstruction of the signal, a de-emphases process is applied to reverse the effects of pre-emphasis. It turns out that in the spectrum for voiced segments like vowels, there is more energy at the lower frequencies, which is called spectral tilt, is caused by the nature of the glottal pulse [31]. Therefore, boosting the high frequency energy makes information from these higher formants more available to the acoustic model and improves phone detection accuracy. Without pre-emphasis, the linear prediction would incorrectly focus on the lower-frequency components of speech, losing important information about certain speech segments. The pre-emphasis phase is performed by a first-order high-pass filter. In the time domain, with input $x(n)$ and 0.9⩽α⩽1.0, the filter equation is $y[n] = x[n] - \alpha x[n-1]$. Since the spectrum of a vowel spoken by an average human being falls off with approximately 6 dB per octave, a pre-emphasis frequency of 50 Hz is applied in this work.

Deconvolution based on linear predictive analysis

Given a convolved signal $y(t) = x(t) * h(t)$, deconvolution is applied to isolate the components $x(t)$ and $h(t)$. Several deconvolution methods have been proposed for speech analysis, such as cepstral analysis and linear predictive coding [32]. The main disadvantage of cepstral analysis is that the speech deconvolution is performed in the frequency domain, which leads to a certain amount of computational complexity. Therefore, the linear prediction analysis is applied in this work as it can be performed in the time domain. According to linear predictive coding, the *n*th sample in a set of speech samples can be predicted by the weighted amount of the p previous samples illustrated as:

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{4.5}$$

where a_k denotes the weights on the previous samples, and the number of samples p denotes the order of the LPC during the speech analysis phase. The optimal values of weights a_k are selected to minimise the mean error $e(n)$ between the genuine speech samples and its predicted value.

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{4.6}$$

After applying the *z*-transform, the error signal $E(z)$ can be represented as the multiplication of the speech signal $S(z)$ and the transfer function $A(z)$.

$$E(z) = S(z)\left[1 - \sum_{k=1}^{p} a_k z^{-k}\right] = S(z)A(z) \tag{4.7}$$

Therefore, the speech signal $S(z)$ can be represented as:

$$S(z) = \frac{E(z)}{A(z)} \tag{4.8}$$

where the transfer function $1/A(z)$ represents an all-pole digital filter, and the coefficients a_k correspond to the poles in the filter's z-plane.

4.2.3 Filter/Spectral de-identification

By filter de-identification, we mean the modification of the magnitude spectrum of the frequency response of the vocal tract system (Stylianou 2009). After the determination of LPC coefficients a_k, the filter $A(z)$ can be obtained following the Equation 4.15. The excitation component $e(n)$ can be obtained by filtering the original signal $s(n)$ with the inversed filter $1/A(z)$. Finally, the filter $A(z)$ is replaced by a new filter extracted from a different speaker, while the excitation component remains the one from the source speaker. The procedure of filter de-identification is illustrated in Figure 4.2.

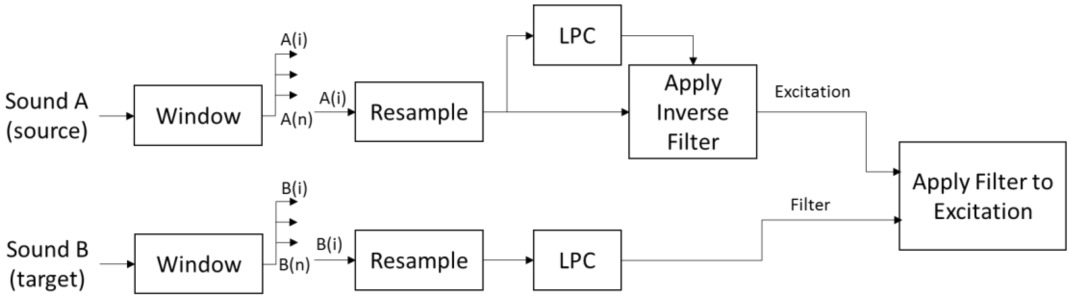

Figure 4.2 Procedure of the filter de-identification

4.2.4 Source/Prosodic de-identification

Apart from the filter de-identification, the source de-identification phase mainly includes three types of modification: pitch modification, time-scale modification, and energy modification [33].

Pitch modification

The goal of pitch modification is to alter the fundamental frequency in order to compress or to expand the spacing between the harmonic components in the spectrum while preserving the short-time spectral envelope as well as the time evolution. It has to be clarified that the pitch and fundamental frequency are used interchangeably in this context though they differ in their definitions. However, pitch is in fact the perceived fundamental frequency. Pitch-Synchronous Overlap and Add (PSOLA) is a method used to manipulate the pitch of a speech signal to match it to that of the target speaker [34]. The basic algorithm for the PSOLA technique consists of three steps:

- ❐ Firstly, the speech signal is divided into separate but overlapping signal segments. This is accomplished by windowing segments around each pitch mark or peak amplitude in the original signal.
- ❐ Secondly, the separated signals are modified by either repeating or leaving out speech segments, depending on whether the pitch of the target speaker is higher or lower than the pitch of the source speaker.
- ❐ Finally, the remaining signal segments are recombined via the method of overlapping and adding. The re-synthesised signal contains different fundamental frequency compared to the source speaker, while remains the same spectrum characteristics.

Duration modification

The duration-scale de-identification aims to change the speaker duration characteristics through modifying the speaking rate of articulation. This means that the formant structure of the input speech is changed at a slower or faster rate than the rate of the input speech, otherwise the structure is not de-identified [33]. The appropriate time-scale de-identification factor $\beta(t)$ can be estimated as:

$$\beta(t) = \sum_{i=1}^{L} v_i \frac{d_i^t}{d_i^s} \tag{4.9}$$

where i denotes index of signal frame, L denotes the total number of signal frames, v_i represents the value of weights, and d_i^s and d_i^t represent average duration statistics of source speaker and de-identified speaker respectively.

Energy modification

In addition to pitch and duration, energy is another important component which characterises the prosody of a speaker. The goal of energy de-identification is to modify the perceived loudness of the input speech [33]. It is considered to be the simplest de-identification among the prosodic de-identification since the signal is just multiplied by a scale factor which corresponds in amplifying or attenuating all the frequencies by the same factor. The signal energy is scaled with a variable γ at each frame. The scaling factor can be expressed as follows:

$$\gamma(t) = \sum_{i=1}^{L} w_i \frac{e_i^t}{e_i^s} \tag{4.10}$$

where i denotes index of signal frame, L denotes the total number of signal frames, w_i represents the value of weights, and e_i^s and e_i^t denote average energy characteristics of source speaker and de-identified speaker, respectively.

4.3 Speaker verification experiments

According to the evaluation criteria defined in Chapter 2, the de-identification performance of the described speaker de-identification systems can be rated by the state-of-the-art speaker recognition system.

4.3.1 Speaker verification system

In this work, the conventional GMM-UBM speaker recognition techniques is applied [35]. The de-identification performance against speaker identification systems is evaluated by calculating de-identification score DIR for all experiments as defined in Chapter 2.

In the conventional GMM-UBM framework, the universal background model (UBM) is a Gaussian mixture model (GMM) that is trained on a pool of background data from a large number of speakers [36]. The speaker models are then adapted from the UBM using the maximum a posteriori (MAP) estimation. During the evaluation phase, each test utterance is scored against the background model and the given speaker model to accept or reject an identity claim. Here, the number of Gaussian mixture component used is 64 and the GMMs for each speaker in the database are trained using 10 iterations of EM algorithm.

4.3.2 Speaker de-identification systems

For baseline systems using GMM and WFW, the analysis rate of speech files is 128 samples per frame and the Hamming window is used for windowing. We set the pitch range from 60 to 500 Hz and maximum voiced frequency considered is 5500 Hz. In voice conversion, spectral parameterizations using line spectral frequencies (LSFs) with 14th order LPC filter is considered due to its good interpolation properties. On the other hand, as MFCC has been used extensively for speaker recognition research, we use MFCC as the feature for all speaker recognition experiments. Here, 39 dimensional MFCCs including delta and delta–delta coefficients are extracted using linearly spaced filters in Mel scale [37]. Here, features are not further processed for RASTA filtering, voice activity detection (VAD), utterance level cepstral mean and variance normalization (CMVN), because it will not affect our final aim i.e. to compare the de-identification performance of different speaker de-identification systems against the speaker recognition systems. For both the GMM and WFW based approaches, we set the order of GMM models as 8 for the voice conversion experiments.

For the new speaker de-identification solution based on LPC, the pre-emphasis frequency of 50 Hz is applied to all the speech signals in the database. Then due to the 5500-Hz band-limiting implementation, we resample all the segments to 11 kHz. For the linear-prediction analysis on the resampled frames, the LPC filter of 28th order is applied in this work. During the pitch analysis phase, the scope of pitch analysis is between 75 hertz and 600 hertz.

4.3.3 Database design

CMU ARCTIC [38] is a parallel database provided by the Carnegie Mellon University, which contains 7 different speakers with 540 parallel utterances from each of them [38]. The seven speakers in the database are coded as: BDL (US English, US male), SLT (US English, US female), CLB (US English, US female), RMS (US English, US male), JMK (US English, Canadian male), AWB (US English, Scottish male) and KSP (US English, Indian male). It provides 16-bit, 16 kHz sampled speech waveform file for each utterance and most of the recorded sentences are between 1 and 4 s duration.

Due to the fact that a large number of utterances are included in the CMU ARCTIC database, we use this corpus for target speaker model training and testing in the speaker verification experiments. To perform this task, 3 male (BDL, RMS, KSP) and 2 female (CLB, SLT) speakers are taken from the corpus. It is worth noting that we have not used parallel sentences of each speaker in the experiments because they may introduce an unfair bias for certain phonemes in the speaker recognition procedure. This is why we use 30 utterances of each speaker for target speaker model training and another 30 utterances for testing. This forms baseline corpus in these speaker verification experiments.

Additionally, 100 utterances of each 5 speakers is utilised to conduct genuine experiments to set the decision threshold of the speaker verification system. By genuine experiment, we mean no de-identification is applied here. In particular, the fixed decision threshold will be applied for the speaker verification system to decide whether the de-identified data are to be accepted or rejected for a fixed target speaker. Thus, the number of genuine target trials will be 500 (5×100) and the number of genuine imposter trials will be 2000 (5×100×4).

We then design the testing corpus for the three described speaker de-identification systems (GMM, WFW, LPC) described in Section 4.1 and 4.2. Here, speaker models are the same as the baseline corpus, but the test utterances of each speaker are de-identified through three speaker de-identification techniques. Speaker de-identification is carried out through all possible transformation directions, including M→M, M→F, F→F, and F→M, where M and F denote male and female respectively. Therefore, the number of genuine trials will be 150 (30 test sentences for each 5 speakers). The 30 test utterances of each 5 speakers undergo speaker de-identification to develop the testing corpus. As a result, the number of de-identified testing utterances will be 600 (5×30×4).

TIMIT [39] is an acoustic/phonetic continuous read speech database, which contains broadband recordings of 630 speakers of eight major dialects of American English, each speaker reading ten sentences. The database includes time-aligned orthographic, phonetic and word transcriptions. Due to the large number of speakers, the TIMIT corpus is used for Universal Background Model (UBM) training in the speaker verification experiments. For designing the UBM corpus, all 10 utterances of each 462 speakers from the training section of the database are used. The detailed database design and several statistics regarding all the related experiments are illustrated in Table 4.1.

Table 4.1 Statistics regarding speaker verification experiments

	Male	Female	Total
Target speakers	3	2	5
No of speakers in UBM	326	136	462
Genuine target trials	300	200	500
Genuine imposter trials	1200	800	2000
Converted test trials	360	240	600

4.3.4 Experiment results and discussion

As described in Section 4.3.3, genuine experiments are conducted to set the decision threshold of the speaker verification system to decide whether the claim is to be accepted or rejected for a fixed target speaker. The total number of test utterances here is 500, and the genuine test results are shown in Table 4.2. The decision threshold is determined when the FAR is 0% for 500 target trials and the FRR is 0% for 2000 imposter trials. This indicates an excellent recognition performance of 100% against the fixed threshold.

The experiment results using the GMM-UBM speaker verification system are given in Table 4.3. Here, we report de-identification rate (DIR) values using all three speaker de-identification techniques for all possible conversion directions. The larger the DIR, the better de-identification performance of the speaker de-identification system. From Tables 4.3, we can see that for three speaker de-identification techniques (GMM, WFW and LPC), the DIR score of the LPC based system is much higher than the GMM based and WFW based systems. Thus, it can be said that the LPC based speaker de-identification method produces more desired performance on speaker de-identification than GMM based and WFW based approaches. The main reason for the higher de-identification score of LPC based method is that it captures overall features in both spectral and prosodic domains. Whereas, the GMM de-identifies only trends in spectral power and WFW is basically a hybrid approach between GMM and DFW techniques that tries to do a compromise between capturing overall trends in spectral power and maintaining spectral details by using an energy correction filter. Here, during frequency warping phase, a certain amount of spectral contents remains intact as the source speaker, which degrades the de-identification performance of the transformed speech in WFW based method. Although a 'perfect' de-identification performance can be seen using the LPC-based new strategy from Table 4.2, it can be also caused by the speech quality degradation in the de-identified speech. In order to prove the intelligibility of the de-identified speech, the speech recognition experiments will be conducted as the next step.

Table 4.2 Speaker verification results using genuine utterances

SV system used	Test utterances	Target trials	Imposter trials	FAR (%)	FRR (%)
GMM-UBM	500	500	2000	0	0

Table 4.3 Speaker verification results using three speaker de-identification systems

SDI technique used	Test utterances	Accepted	Rejected	DIR (%)
GMM	600	84	516	86.0000
WFW	600	157	443	73.8333
LPC	600	0	600	100

5. Future work plan

According to the scope of work described in Section 1.3, the research project can be divided into 5 work packages as:

- ❒ Work 1: Risk assessment against speaker identification
- ❒ Work 2: Literature review for speaker de-identification solutions
- ❒ Work 3: Development of high-quality speaker de-identification strategy
- ❒ Work 4: Evaluation for speaker de-identification methodologies
- ❒ Work 5: Integration of speaker de-identification system into voice user interface

During the last one year, initial efforts have been made to complete the Work 1 and Work 2 packages, including the risk assessment and the literature review. Additionally, Work 3 package has been partially completed, including the implementation of the state-of-the-art baseline speaker de-identification systems and the development of one novel speaker de-identification strategy as described in Chapter 4.

Thus, the future works will continue to focus on Work 3 to Work 5 packages during the next period of 24 months. The thesis is to be completed during the last six months till September in 2018. The work packages, individual tasks, deliverables and milestones are shown in the following Gantt Chart Table 5.1. Detailed tasks within each work package are described as follows:

Work 3: Development of high-quality speaker de-identification strategy

- ❒ Task 3.1: Efficiency-enhanced speaker de-identification strategy. The objective of this task is to enhance the efficiency of the speaker de-identification performance. One of the solutions is to combine both the prosodic de-identification and spectrum de-identification, which theoretically preserves better de-identification performance.
- ❒ Task 3.2: Intelligibility-preserved speaker de-identification strategy. The goal in this task is to preserve the intelligibility of the de-identified speech signals. One of the possible steps to achieve better intelligent speech is to use a promising approach of speech synthesis, which is based on the deep-belief networks. Moreover, it is worth exploring new target speaker selection methods to reconstruct more intelligent sounding speech.

- Task 3.3: Non-Limitation on target speakers to be de-identified. The aim here is to allow any unseen new users can use the speaker de-identification system for speaker de-identification, without providing speech samples for feature training and enrolling at the system in advance. This aim can be accomplished by developing an online speaker modification system.
- Task 3.4: Reversibility- and non-reversibility-assured speaker de-identification strategy. This task aims to allow authorised listeners to retrieve the original identity of the source speaker but to avoid un-authorised listeners re-identifying the source speaker. One of the solutions is to transform a speaker key that would allow authorised user to transform the de-identified speech back to the original speech but to a certain extent, avoiding unauthorised user to re-identify the speech.
- Task 3.5: Individuality-guaranteed speaker de-identification strategy. The objective of this task is to guarantee individuality between different speakers. One of the possibilities to achieve this goal is to apply different acoustic transformation methods to the speech provided by different speakers.

Work 4: Evaluation for speaker de-identification methodologies

- Task 4.1: Efficiency evaluation for speaker de-identification approaches. This objective evaluation experiment will be performed by the state-of-the-art speaker identification systems and speaker verification systems.
- Task 4.2: Intelligibility evaluation for speaker de-identification approaches. The intelligibility of de-identified speech will be evaluated by means of the state-of-the-art speech recognition systems.
- Task 4.3: Subjective evaluation for speaker de-identification approaches. The efficiency of speaker de-identification and the intelligibility of the de-identified speech will be also rated by real human listeners, so that the final performance scores are reliable and give an idea of the impact that the resulting system can have in real world.

Work 5: Integration of speaker de-identification system into the voice user interface

- Task 5.1: Implementation of the voice-driven user interface system.
- Task 5.2: Integration of the resulting speaker de-identification system into a voice UI system.

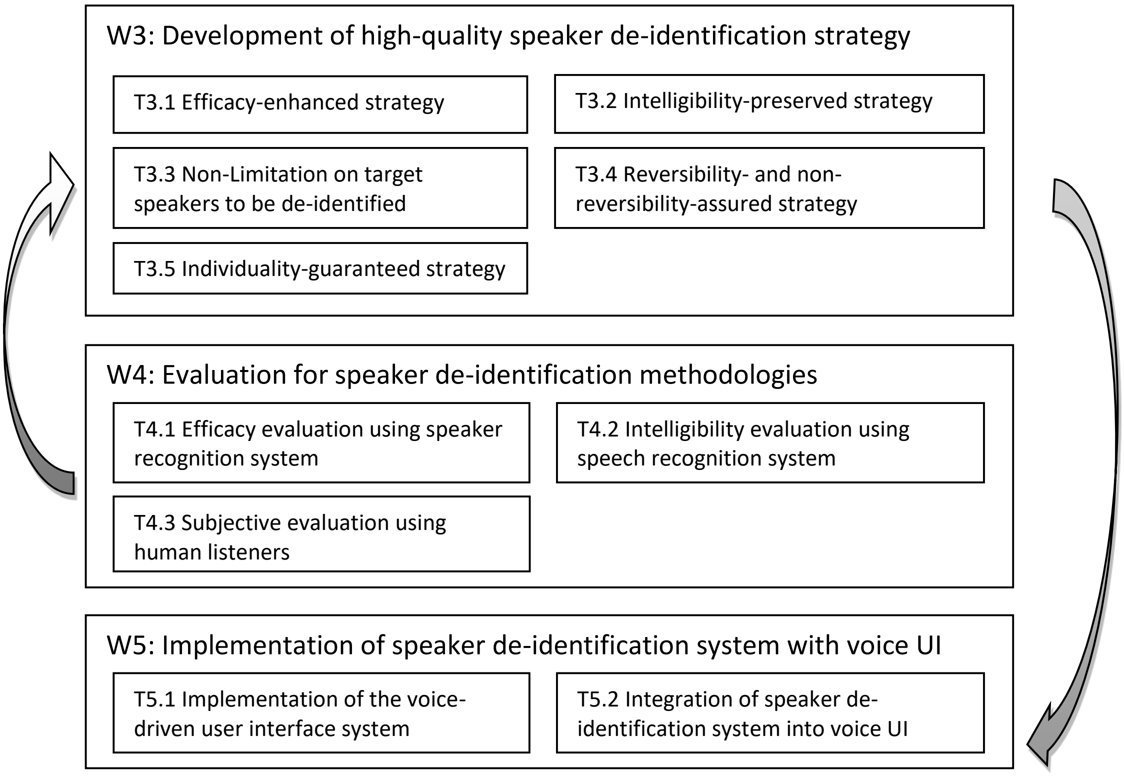
W3: Development of high-quality speaker de-identification strategy
T3.1 Efficacy-enhanced strategy
T3.2 Intelligibility-preserved strategy
T3.3 Non-Limitation on target speakers to be de-identified
T3.4 Reversibility- and non-reversibility-assured strategy
T3.5 Individuality-guaranteed strategy
W4: Evaluation for speaker de-identification methodologies
T4.1 Efficacy evaluation using speaker recognition system
T4.2 Intelligibility evaluation using speech recognition system
T4.3 Subjective evaluation using human listeners
W5: Implementation of speaker de-identification system with voice UI
T5.1 Implementation of the voice-driven user interface system
T5.2 Integration of speaker de-identification system into voice UI

Figure 5.1 Framework of future work plans

Table 5.1. Gantt chart

Tasks (T) and Deliverables (D)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
W1: Risk assessment against SDI	🟦	🟦																																		
W2: Literature review for SDI strategies		🟦	🟦	🟦	🟦	🟦																														
D1: Literature survey						1																														
W3: High-quality SDI solution							🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦																
T3.1 Efficacy-enhanced strategy							🟩	🟩	🟩	🟩	🟩	🟩	🟩																							
T3.2 Intelligibility-preserved strategy													🟩	🟩	🟩	🟩	🟩	🟩																		
T3.3 Non-Limitation on speakers									🟩	🟩	🟩	🟩																								
T3.4 (Non-)Reversibility-assured strategy																	🟩	🟩	🟩	🟩																
T3.5 Individuality-guaranteed strategy											🟩	🟩	🟩	🟩																						
D2: Year 1 report												2																								
W4: Evaluation for SDI methods										🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦													
T4.1 Efficacy evaluation										🟩	🟩	🟩	🟩	🟩	🟩																					
T4.2 Intelligibility evaluation														🟩	🟩	🟩	🟩	🟩	🟩																	
T4.3 Subjective evaluation																			🟩	🟩	🟩	🟩	🟩													
D3: Year 2 report																								3												
W5: Integration of SDI system into voice UI																							🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦						
T5.1 Implementation of voice UI																							🟩	🟩	🟩	🟩	🟩									
T5.2 Integration of SDI into voice UI																										🟩	🟩	🟩	🟩	🟩						
PhD thesis writing																															🟦	🟦	🟦	🟦	🟦	🟦
D4: PhD thesis																																				4

6. Conclusions

In this report we systematically review speaker de-identification solutions that have been proposed in the literature. We notice that most schemes are based on voice conversion techniques, modifying the voice produced by a specific speaker, for it to be perceived by listeners as if it had been uttered by a different speaker. In the following text we discuss our main findings. First, the efforts of researchers have been focused mainly on de-identifying acoustic spectral features of speech; however, de-identifying prosodic characteristics of voice is still an important challenge. Second, currently no scheme exists that achieves good scores on both de-identification and intelligibility. Third, state-of-the-art schemes for speaker de-identification require system training and previous enrolment of speakers which brings limitations in real-world applications. Fourth, no existing scheme is capable of assuring both reversibility and non-reversibility of the de-identified speech. Fifth, it is not possible to distinguish different speakers after de-identification at the present. Sixth, the schemes need to be evaluated further using larger speaker databases in order to verify the accuracy of evaluation experiments.

Furthermore, in this report we proposed a novel solution to speaker de-identification, which relies on the combination of spectral de-identification and prosodic de-identification techniques. This new strategy addresses the first and third problems as described in the former paragraph. In particular, this method does not require system training or previous enrolment of speakers for de-identification, thus significantly extending possible applications in real-world scenarios. We evaluated the proposed method through objective speaker verification experiments to assess the de-identification performance. In order to prove the validity of the new method, it is compared to the state-of-the-art systems based on statistical Gaussian mixture model. The experimental results indicate that the proposed method produces more desired performance on speaker de-identification than the state-of-the-art baseline systems. The main reason for the higher de-identification score of the new method is that it captures overall identifiable voice features in both spectral and prosodic domains.

Research for the next two years will continue to focus on high-quality speaker de-identification strategy. This high-quality method aims to address the rest of the problems described in the first paragraph (e.g. second, fourth, fifth and sixth problems). In addition, speech recognition experiments will be conducted to examine the speech intelligibility performance. The resulting speaker de-identification system will be integrated into a real-life voice-driven application.

References

[1] T. Justin, V. Struc, S. Dobriˇ, B. Vesnicer, I. Ipšić, and F. Miheliˇ, “Speaker de-identification using diphone recognition and speech synthesis,” in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2015, vol. 4, pp. 1–7.

[2] Z. Wu and H. Li, “Voice conversion versus speaker verification: an overview,” in *APSIPA Transactions on Signal and Information Processing*, 2014, vol. 3, p. e17.

[3] I. Alice, “Biometric recognition: security and privacy concerns,” *IEEE Secur. Priv.*, 2003.

[4] J. A. Markowitz, “Voice biometrics,” *Commun. ACM*, vol. 43(9), pp. 66–73, 2000.

[5] Wikipedia contributors, “Speaker recognition,” *Wikipedia, The Free Encyclopedia*, 2016. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Speaker_recognition&oldid=714433656.

[6] M. Pal and G. Saha, “On robustness of speech based biometric systems against voice conversion attack,” *Appl. Soft Comput.*, vol. 30, pp. 214–228, 2015.

[7] Q. Jin and A. R. Toth, “Voice convergin: speaker de-identification via voice transformation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3909–3912.

[8] M. Farrús, D. Erro, and J. Hernando, “Speaker recognition robustness to voice conversion,” *Jornadas Reconoc. Biométrico Pers.*, pp. 73–82, 2008.

[9] T. S. Qin Jin, Arthur R. Toth, Alan W Black, “Is voice transformation a threat to speaker identification?,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4845–4848.

[10] Q. Jin, A. R. Toth, T. Schultz, and A. W. Black, “Speaker de-identification via voice transformation,” in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2009, pp. 529–533.

[11] M. Pobar and I. Ipši´, “Online speaker de-identification using voice transformation,” in *Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2014, pp. 1264–1267.

[12] M. Abou-zleikha, Z. Tan, and M. Græsbøll, “A discriminative approach for speaker selection in speaker de-identification systems,” in *IEEE Signal Processing Conference (EUSIPCO)*, 2015, pp. 2102–2106.

[13] A. Kain and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, vol. 1, pp. 285–288.

[14] D. E. Eslava and A. M. Bilbao, “Intra-lingual and cross-lingual voice conversion using harmonic plus stochastic models,” Barcelona, Spain: PhD Thesis, Universitat Politechnica de Catalunya, 2008.

[15] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, “The HTK book (v3. 4),” 2006.

[16] J. Gros, N. Pavešić, and F. Mihelič, “Text-to-Speech Synthesis: A Complete System for the Slovenian Language,” *CIT. J. Comput. Inf. Technol.*, pp. 11–19, 1997.

[17] D. R. Reddy, “Speech recognition by machine: a review,” *Proc. IEEE*, vol. 64, no. 4, pp. 501–531, 1976.

[18] K. F. Lee, H. W. Hon, and R. Reddy, “An overview of the SPHINX speech recognition system,” *IEEE Trans. Acoust.*, vol. 38, no. 1, pp. 35–45, 1990.

[19] P. C. Woodland, M. J. F. Gales, D. Pye, and S. J. Young, “The development of the 1996 HTK broadcast news transcription system,” in *DARPA speech recognition workshop*, 1997.

[20] Sun/Oracle, “Java Speech API,” 1998. [Online]. Available: http://www.oracle.com/technetwork/java/jsapifaq-135248.html.

[21] Microsoft, “Microsoft Speech Application SDK,” 2005. [Online]. Available: https://msdn.microsoft.com/en-us/library/ms986944.aspx.

[22] I. McGraw, R. Prabhavalkar, R. Alvarez, M. G. Arenas, K. Rao, D. Rybach, O. Alsharif, H. Sak, A. Gruenstein, F. Beaufays, and C. Parada, “Personalized Speech recognition on mobile devices,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5955–5959.

[23] D. P. John Garofolo, David Graff, Doug Paul, “CSR-I (WSJ0) Complete,” *Philadelphia: Linguistic Data Consortium*, 2007. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S6A.

[24] S. Martincid-ipsic and I. Ipsic, “VEPRAD: a Croatian speech database of weather forecasts,” in *Proceedings of the 25th International Conference on Information Technology Interfaces*, 2003, pp. 321–326.

[25] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 131–142, 1998.

[26] D. Erro and A. Moreno, “Weighted frequency warping for voice conversion,” in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2007, vol. 2, pp. 1965–1968.

[27] Y. Stylianou, “Applying the harmonic plus noise model in concatenative speech synthesis,” *IEEE Trans. Speech Audio Process.*, vol. 9, no. 1, pp. 21–29, 2001.

[28] D. Erro, A. Moreno, and A. Bonafonte, “Flexible harmonic/stochastic speech synthesis,” in *6th ISCA Workshop on Speech Synthesis (SSW)*, 2007, pp. 194–199.

[29] D. Erro, A. Moreno, and A. Bonafonte, “INCA algorithm for training voice conversion systems from nonparallel corpora,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 5, pp. 944–953, 2010.

[30] L. M. Arslan, “Speaker transformation algorithm using segmental codebooks (STASC),” *Speech Commun.*, vol. 28, no. 3, pp. 211–226, 1999.

[31] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice Hall, 2009.

[32] R. Lawrence and S. Ronald, *Digital processing of speech signals*. Prentice Hall, 1978.

[33] Y. Stylianou, “Voice transformation: A survey,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3585–3588.

[34] H. Valbret, E. Moulines, and J. P. Tubach, “Voice transformation using PSOLA technique,” *IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 1, pp. 145–148, 1992.

[35] S. Omid Sadjadi, M. Slaney, and L. Heck, “Msr identity toolbox v1. 0: a matlab toolbox for speaker-recognition research,” 2013.

[36] D. a. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digit. Signal Process.*, vol. 10, no. 1, pp. 19–41, 2000.

[37] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: from features to supervectors,” *Speech Commun.*, vol. 52, no. 1, pp. 12–40, 2010.

[38] “CMU_ARCTIC speech synthesis database,” *Carnegie Mellon University*, 2003. [Online]. Available: http://festvox.org/cmu_arctic/index.html.

[39] J. Garofolo, “TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1,” *Philadelphia: Linguistic Data Consortium*, 1993. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S1.