# A DISCRIMINATIVE APPROACH FOR SPEAKER SELECTION IN SPEAKER DE-IDENTIFICATION SYSTEMS

*Mohamed Abou-Zleikha[1,2], Zheng-Hua Tan[2], Mads Græsbøll Christensen[1], Søren Holdt Jensen[2]*

[1]Audio Analysis Lab, AD:MT, Aalborg University, Denmark
[2]Department of Electronic Systems, Aalborg University, Denmark
{moa,mgc}@create.aau.dk,{moa,zt,shj}@es.aau.dk

## ABSTRACT

Speaker de-identification is an interesting and newly investigated task in speech processing. In the current implementations, this task is based on transforming one speaker speech to another speaker in order to hide the speaker identity. In this paper we present a discriminative approach for human speaker selection for speaker de-identification. We used two modules, a speaker identification system and a speaker transformation one, to select the most appropriate speaker to transform the source speaker speech from a set of speakers. In order to select the target speaker, we minimize the identification confidence of the transformed speech as the source speaker and maximize the confusion about the transformed speech membership to the rest of the speaker models and the identification confidence of the re-transformed speech using the source speaker model. These three factors are combined to achieve overall optimization performance in order to select the best target speaker to transform the source.

***Index Terms***— speaker de-identification, speaker identification, speaker transformation

## 1. INTRODUCTION

Information privacy is considered a very important factor in the media technology. Nowadays, several applications consider the speaker identity- not the content -a confidential information and require hiding the speaker identity. Some examples include training to employees in calling centers and listening to the recorded medical reports where the students or employees are allowed to know the content but not the speaker. Other applications require hiding the speaker identity for security reasons, such as some radio program interviews and court witnesses. Another type of applications, such as telephone banking services, require to transmit the speaker speech without revealing the speaker identity during the transmission and then back-transform the speaker to its original for the authorized listeners. Such application requires both a speaker de-identification and a re-identification. All these applications require hiding the speaker identity. This speaker identity hiding task is called speaker de-identification. A perfect speaker de-identification system meets two requirements: (1) it does hide the speaker's true identity to any unauthorized listener (referred to as speaker de-identification), and (2) it transmits a key which allows the authorized listeners to retrieve the original identity of the speaker from the transformed speaker (speaker re-identification) [1].

Few attempts have been made to build de-identification systems [2], [1], [3]. The earliest attempt for speaker de-identification started from the idea if speaker transformation can deceive the speaker identification system [2]. In that study, the authors used diphone-based syntactic speech (kal-diphone) as a source and they transformed it to a set of speakers attempting to fool the speaker identification system. The results showed that the speaker transformation could fool the purely acoustic speaker identification system (Gaussian Mixture Model (GMM) based one). In later research, they proposed the usage of speaker transformation to carry out speaker de-identification as a part of a secure speech transmission system [4, 1]. In this research, the authors studied several transformation techniques where the target speech was still the kal-diphone syntactic speech.

In [3], speaker transformation was used to de-identify a speaker whose speech has not been used to build (i.e. to train) speaker transformations. They used a syntactic speech as a target speaker as well (voice generated using a Hidden Markov Model (HMM) speech synthesiser). They claimed -as in other previous work- that using the syntactic speaker has a certain amount of vocoded buzzy character that was also present in the target speaker speech. Using the syntactic speaker decreased the naturalness of the de-identified speech and the performance of the de-identification system. All the existing work aimed at:

- getting lowest identification confidence to be identified as the source and
- hoping to obtain the original speaker as good as possible if we reverse the de-identification operation in order

This work was supported in part by the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation under the CoSound project, case number 11-115328. This publication only reflects the authors' views.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

to do the re-identification.

However, when the target speaker is chosen to be syntactic, the two goals can hardly be achieved.

This paper proposes a target speaker selection method to choose a speaker from a repository of human speakers to transform to in speaker de-identification system. The proposed method aims at selecting a speaker that:

- gives the lowest identification confidence to be identified as the source,
- does not converge to a certain speaker completely, but gives as much doubt as possible about the speaker identity, and
- can give a good result if we reverse the de-identification operation in order to do the re-identification.

The proposed method follows a discriminative approach that minimises the identification confidence of the transformed speaker to be identified as the source and maximizes the ambiguity about the speaker identity and the source identification confidence when we back-transform the transformed speech to the original. The method provides a closed loop of a secure speech transmission task that gives everybody access to what was said but gives the information about who said it to the authorised people only.

This paper is organized as follows: in Section 2, the baseline speaker de-identification components using speaker transformation is presented. In Section 3 the proposed approach for speaker selection is explained. In Section 4, experimental results on the CHAINS corpus are provided, and Section 5 presents the conclusion and future work.

## 2. SPEAKER DE-IDENTIFICATION USING SPEAKER TRANSFORMATION

### 2.1. Speaker Transformation

Speaker transformation attempts to transform speech from the source speaker to the target speaker, so the transformed speech sounds as if it was produced by the target. In this paper, the weighted frequency warping for speaker transformation has been used [5]. The weighted frequency warping is a spectral envelope conversion method based on time-varying frequency warping transformations combined with GMMs. This combination brings together the advantages of both approaches. The two reasons behind choosing this system for speaker transformation are as follows. The first reason is that there is an open source toolbox [6]. The second reason is that this speaker transformation system showed very good and balanced results between the output quality Mean Opinion Score (MOS) and the transformation similarity MOS (which is define the similarity between the source voice and the transformed voice) as shown in [6], which is preferred to the toolboxes that give very high MOS for quality but a high MOS for similarity at same time or low MOS for similarity but a low MOS for quality [6].

### 2.2. Speaker Identification

For the speaker identification component, two approaches have been used: GMM-based and i-vector-based speaker identification. The GMM-based speaker identification is based on training a universal background model (UBM) represented by a GMM using all speakers available. Then, a set of speaker-specific models are generated by adapting the UBM using the maximum a posteriori (MAP) estimation [7]. During the evaluation phase, each test segment is scored against all trained speaker models, and the speaker model that has the highest score is the identified one.

The i-vector speaker identification approach is the most popular and is nowadays considered state-of-the-art [8, 9]. Under this approach, a UBM is built as in the GMM-based approach. Then, the total variability subspace from background data is extracted, the development i-vectors are extracted and their dimensions are reduced using Linear Discriminant Analysis (LDA). The Gaussian probabilistic LDA (PLDA) models with dimension reduced i-vectors are trained for each speaker (the same of the training data could be use for training the Gaussian PLDA if the size of the data is small). During the evaluation phase, each test segment is scored against all trained speaker models using the dimension-reduced i-vectors of the testing data, and the speaker model that has the highest score is the identified one. We refer to the normalized scoring values of each speech segment for all speaker models as the identification confidence.

## 3. THE PROPOSED APPROACH FOR SPEAKER SELECTION FOR THE SPEAKER DE-IDENTIFICATION

The proposed approach for speaker selection starts by calculating the transformation models between the source speaker and the speakers in the repository and vice versa. In addition, a speaker identification system is built using the speakers in the repository and the source speaker. These two components are used to calculate the scores for selecting the ideal speaker from the repository.

In order to select the ideal speaker, the three goals introduced in Section 1 are required to be achieved. Firstly, to achieve the lowest identification confidence to be identified as the source, we calculate the identification confidence for the transformed speech using the speaker identification system and try to minimise the identification confidence value when using the source speaker model. Then, the confusion requirement is achieved by calculating the confusion about the membership of the transformed speech to any of the speakers in the speaker repository. And later, the confusion factor is

defined as the uncertainty degree of the identification over all speakers. By maximizing this confusion factor, we increase the doubt about the speaker identity so the converted speech would not be recognised as any of the repository speakers.

These steps work well for hiding the speaker identity. However, they also make retrieving the original speaker difficult. To solve this issue, we e-transform the transformed speech back to the original speaker, then calculate the identification confidence for it and try to maximize the speaker identification system for the original speaker.

Each of the previously mentioned calculations tries to optimize one of the goals introduced in Section 1, leading to optimal de-identification performance for each particular speaker. In order to achieve overall optimisation performance, we combine these three functions into one as

$$k_i = \arg\max_{k \in K}(-\alpha f(i,k) + \beta c(i,k) + \gamma d(i,k)), \quad (1)$$

where $f(i,k)$ is the identification confidence of the transformed speech to the original speaker, $c(i,k)$ is the confusion factor of the transformed speech from the source $i$ to the speaker $k$, $d(i,k)$ is the identification confidence of the re-transformed speech to the original speaker, $k$ is the speaker index, $K$ is the number of speakers in the repository, and $\alpha$, $\beta$ and $\gamma$ are the weights of the previously mentioned functions respectively. The first function is multiplied by $-1$ because we want to minimize this term. Figure 1 illustrates the general framework of the speaker selection process. The process starts by having the source speaker and a repository of speakers. These two inputs are used to build a speaker identification system and a a set of a transformation models between the source and each speaker in the speaker repository. Using these two modules, the target speaker is selected by applying the equation 1, obtaining a transforming model (which allows to transform the source speaker to the target speaker) and the speaker key, which allows to re-transform the transformed speech back to the source speaker.

### 3.1. Confusion Factor

To calculate the confusion factor of a transformed speaker, we consider the impurity of the identification confidence of the transformed speaker using the repository speakers. To calculate the impurity, we used two measures, the entropy measure and the Gini index, respectively.

#### *3.1.1. Entropy*

The confusion factor for a speaker $i$ transformed to a speaker $k$ using the entropy is calculated as:

$$c(i,k) = -\sum_{j=1}^{N} p_j log(p_j) \quad (2)$$

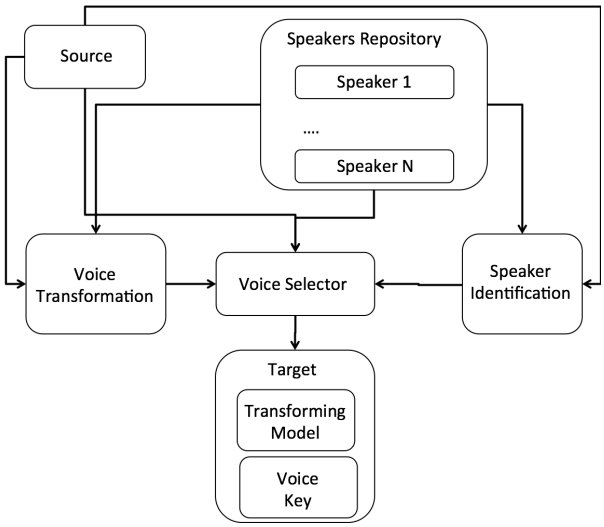


**Fig. 1**. General framework of the speaker selection process

where $N$ is the number of speakers in the repository and $p_j$ is the identification confidence that the transformed data from $i$ to $k$ is recognised as the speaker $j$.

#### *3.1.2. Gini Index*

The confusion factor for a speaker $i$ transformed to a speaker $k$ using the Gini index measure is calculated as:

$$c(i,k) = 1 - \sum_{j=1}^{N} p_j^2 \quad (3)$$

where $N$ is the number of speakers in the repository and $p_j$ is the identification confidence that the transformed data from $i$ to $k$ is recognised as speaker $j$.

Using the previously mentioned impurity measures as part of (1), we select the best speaker to transform to.

## 4. EVALUATION

### 4.1. Database Description

To evaluate the proposed approach for speaker selection, the CHAINS Corpus has been used. CHAINS is a speech corpus collected for helping the speaker identification research. It consists of 36 speakers, 28 of which (14 male, 14 female) are from the Eastern part of Ireland, and speak Eastern Hiberno-English. The remaining 8 speakers (4 male, 4 female) are from the UK and the USA under several conditions. There was a solo condition, where each speaker read the same four short fables and 33 individual sentences without any noise. In this case, we have parallel speech, where all the speakers have read exactly the same text. The reader can find a full corpus specification in [10]. The four fables and 6 sentences have been used for training the speaker transformation and speaker

| | De-identification | Identification |
|---|---|---|
| | *Syntactic voice* | |
| **I-vector** | 0.6473 ± 0.3838 | 0.6248 ±0.3893 |
| **GMM** | 0.6183 ±0.4175 | 0.7719 ±0.3225 |
| | | |
| | *Human speaker Selection Approach* | |
| **I-vector+Gini** | 0.8577 ±0.2275 | 0.7049 ±0.3667 |
| **I-vector+Entropy** | 0.8622 ± 0.2115 | 0.7027 ± 0.3653 |
| **GMM+Gini** | 0.9019 ± 0.1877 | 0.8699 ± 0.2509 |
| **GMM+Entropy** | 0.9012 ± 0.1874 | 0.8699 ±0.2509 |

**Table 1**. The average and standard deviation of the de-identification accuracy and identification accuracy of the re-transformed speaker for the baseline approach and the proposed approach.

identification systems and the remaining 27 sentences have been used for testing.

### 4.2. System Setup

For speaker transformation, we used the UPC open source MATLAB toolkit for speaker transformation [11]. The number of Gaussian components used for building the speaker transformation is 16. For speaker identification, we used the MSR Identity Toolbox v1.0 [12, 13]. The GMM-based and ivector-based speaker identification systems use 256 Gaussian components. The models are trained on 23 Mel-frequency cepstral coefficients (MFCCs) +log energy. The universal background model for both GMM and i-vector identification models has been built using the Wall Street Journal (WSJ0) corpus [14].

To evaluate the system, we compare the de-identification rate and the identification rate after retransforming between the speakers selected by the proposed approach and a transformation to a syntactic voice as a baseline. We use a diphone-based voice "kal-diphone" to compare with as it has been previously used in some speaker de-identification research [2, 1]. The same configuration used for transforming to the human speakers has been used to transform to the syntactic voice.

### 4.3. Experiments

To evaluate the proposed approach, we compare it with the baseline method with regard of the de-identification accuracy and the identification accuracy for the retransformed speaker to the target as a source speaker. The de-identification accuracy is defined as the percentage of not identifying the transformed speech from the source to the target as a source speaker, and the identification accuracy is defined as the percentage of identifying the re-transformed speech from the target to the source as a source speaker. The parameters $\alpha$, $\beta$ and $\gamma$ take values $0.1$, $0.05$ and $0.5$ respectively. These values have been manually tuned.

| | De-identification | Identification |
|---|---|---|
| **I-vector+Gini** | 0.9989 ± 0.0064 | 0.4571 ±0.3796 |
| **I-vector+Entropy** | 0.9957 ± 0.0163 | 0.4637± 0.3814 |
| **GMM+Gini** | 0.9996 ± 0.0023 | 0.5687 ± 0.3724 |
| **GMM+Entropy** | 0.9954 ± 0.0275 | 0.5711 ±0.3624 |

**Table 2**. The average and standard deviation of the de-identification accuracy and identification accuracy of the re-transformed speaker for the baseline approach and the proposed approach with $\gamma = 0$.

Table 1 shows the average and standard deviation of the de-identification accuracy and identification accuracy of the re-transformed speaker for the baseline approach and the proposed approach using each of the speaker identification system and the confusion measure proposed in the paper. The experiments show that the proposed approach has a significantly better performance than using the synthetic voice, with $p-value \leq 0.05$. The selected speaker has a considerable effect on the accuracy of both de-identification and the identification of the re-transformed speaker using both the GMM and the i-vector identification system. The experiments also show that the i-vector system was more sensitive to the characteristics of the speaker with a lower de-identification accuracy than the GMM model.

To study the effect of adding the identification of the re-transformed speaker accuracy on the de-identification accuracy, we put $\gamma = 0$ and recalculate the accuracies. Table 2 shows the results obtained. We can see that removing the requirement of re-transforming the transformed speech significantly increases the de-identification accuracy. However, the identification accuracy of the re-transformed speech has significantly decreased as well, which makes considering only the de-identification accuracy unsuitable for optimizing the overall performance of a speaker de-identification system.

## 5. CONCLUSIONS AND FUTURE WORK

In this paper, a discriminative approach to human speaker selection for speaker de-identification has been proposed. This approach consists of two components: a speaker identification system and a speaker transformation system. To select the human target speaker, we calculate the identification confidence of the transformed speech to the source speaker model in the speaker identification system, the confusion factor of identification confidence for the rest of the speakers and the identification confidence of the re-transformed speech to the source. The selected speaker minimizes the identification to the source and, at the same time, maximizes the confusion factor and the identification rate of the re-transformed speech to the source.

To evaluate the proposed approach, we compared its performance with that of a baseline system which used a syn-

tactic "kal-diphone" voice to convert to. The evaluation uses two metrics, the de-identification accuracy and the identification accuracy of the re-transformed speaker.

The obtained results show that using a human speaker as target instead of synthetic voice gives higher de-identification and identification accuracies using both i-vector and GMM. However, i-vector shows a lower de-identification accuracy and a lower identification of the re-transformed speaker.

Future extension of this work includes performing a subjective evaluation of the transformed speech with regard to the quality and the speaker identity. In addition, more experiments are required to be conducted to examine the effect of the parameters of the selection functions.

## 6. REFERENCES


[1] Q. Jin, A. Toth, T. Schultz, and A. Black, "Speaker de-identification via voice transformation," in *IEEE Workshop on Automatic Speech Recognition & Understanding, 2009. ASRU 2009*. IEEE, 2009, pp. 529–533.

[2] Q. Jin, A. R Toth, A. Black, and T. Schultz, "Is voice transformation a threat to speaker identification?," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*. IEEE, 2008, pp. 4845–4848.

[3] M. Pobar and I. Ipsic, "Online speaker de-identification using voice transformation," in *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*. IEEE, 2014, pp. 1264–1267.

[4] Q. Jin, A. Toth, T. Schultz, and A. Black, "Voice convergin: Speaker de-identification by voice transformation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009. ICASSP 2009*. IEEE, 2009, pp. 3909–3912.

[5] D. Erro and A. Moreno, "Weighted frequency warping for voice conversion," in *Interspeech*, 2007, pp. 1965–1968.

[6] A. Machado and M. Queiroz, "Voice conversion: A critical survey," in *Proc. Sound and Music Computing (SMC)*, 2010.

[7] D. A Reynolds, TF. Quatieri, and RB. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[8] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 19, no. 4, pp. 788–798, 2011.

[9] O. Plchot et al., "Developing a speaker identification system for the DARPA," *Proceedings of ICASSP 2013*, pp. 6768–6772, 2013.

[10] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The CHAINS speech corpus: CHaracterizing INdividual Speakers," in *Proc of SPECOM*, 2006, pp. 1–6.

[11] "UPC toolkits for voice transformation," http://aholab.ehu.es/users/derro/software.html, Accessed: 2014-09-1.

[12] S. Sadjadi, M. Slaney, and L. Heck, "MSR identity toolbox v1.0: A matlab toolbox for speaker-recognition research," *Speech and Language Processing Technical Committee Newsletter*, November 2013.

[13] "MSR identity toolbox v1.0," http://research.microsoft.com/en-us/downloads/2476c44a-1f63-4fe0-b805-8c2de395bb2c/, Accessed: 2014-09-1.

[14] J. Garofolo and et. al, "CSR-I (WSJ0) complete LDC93S6A. web download.," *Philadelphia: Linguistic Data Consortium*, 1993.