Multiple Frame Rates for Feature Extraction and Reliable Frame Selection at the Decision for Speaker Identification Under Voice Disguise

Swati Prasad, Zheng-Hua Tan and Ramjee Prasad

Electronic Systems, Aalborg University, Aalborg-9220, Denmark
E-mail: {in_sp; zt; prasad}@es.aau.dk

Received 8 April 2016; Accepted 12 May 2016;
Publication 20 June 2016

Abstract¹

Determining the person who spoke a given speech utterance from a group of people is referred to as Speaker Identification. It is used in crime scenes, surveillance and consumer electronic products like smart TV. But it faces poor performance due to a mismatch between the train and the test speech data, that arises because of the adoption of voice disguise. Therefore, this paper studies the effect of three different types of voice disguises, namely, Fast (non-imitative), Synchronous (Imitative) and Repetitive Synchronous Imitation along with the normal speaking from the CHAINS corpus on the speaker identification performance. Finally, a system combining different frame rates for feature extraction and reliable frame selection at the decision level has been proposed. The evaluated system showed an overall better performance than the baseline systems.

Keywords: Robust speaker identification, voice disguise, biometric, frame selection.

¹This paper is an extension of the paper titled, "Reliable frame selection for speaker identification under voice disguise scenario.", which was presented at the WirelessVITAE, Hyderabad, India, Dec. 2015.

1 Introduction

The problem of finding out the speaker of a given speech utterance from a group of M people is referred to as speaker identification. It is also, commonly referred to as biometric based on voice [1, 2]. It finds applications in crime scenes. For example, a suspicious phone recording has been recovered from a crime scene, then the speaker identification system can be used for identifying the person from the suspect's list. It is also used for monitoring the employees in a security conscious organization, and for border control. Speaker identification can also be used in consumer electronics. A typical example can be smart TV. Here, access to TV can only be given to authorized speakers, like family members. Once the TV starts, a customized channel preference can also be provided to the individuals.

Speaker identification system has two major parts: Training/Enrollment and Testing/Recognition and is shown in Figure 1. In Training, individual speaker models were developed from the features extracted from the training speech data provided by each speaker. Speaker models attempt to catch the unique voice characteristics of an individual. During testing, a speech utterance from an unknown speaker is given. The system matches it with the individual speaker models, and the model which gives the highest similarity measure is decided as the speaker of the given speech utterance [3, 4].

Speaker identification faces poor performance when a mismatch between the training and the test speech data occurs. This is referred to as Mismatch problem [5]. A mismatch can occur due to environmental noise, handset or channel variations or because of ageing, throat infection and emotional state of the person [6–9]. A mismatch can also happen, when people modify their voices. It is done, either to be recognized as a target speaker to steal the target's information/resources, or to hide his/her own identity. Voice modifications or voice disguise can be achieved electronically or non-electronically [10]. In electronic voice disguise, electronic devices are used, either to change the voice of a person or to sound like a target speaker. A typical example of this is a voice changer software [11]. In non-electronic voice disguise, speakers deliberately change their voice either to sound like a target or to hide his/her own identity [12]. A survey revealed that, non-electronic voice disguise is more frequently used in crimes than electronic voice disguise because of its ease of production [13]. A lot of research dealing with mismatch due to environmental noise, and handset/channel variation has been done [14–16], but a very few research has been seen dealing with non-electronic voice disguise mismatch problem. Early research showed that the spectrogram

Figure 1 Speaker identification system. (a) Training the speaker model. (b) Testing/Recognition.

based speaker identification's performance, greatly decreased because of the voice disguise [17, 18]. The effects of 10 different types of voice disguises on the Forensic Automatic Speaker Recognition System (FASRS) has been studied in [19]. It has been found that it decreases identification accuracy and the amount of decrease varies with the type of disguise. In [20], different types of prosodic and source related features were tested on the performance of the speaker identification under voice disguise. Features related to the fundamental frequency F0, outperformed other features. In a recent study [21], the effect of voice disguise on suprasegmental features has been studied. It showed low variability within speakers and high variability between speakers. An AM-FM approximation of the input speech signal based feature called

pykfec has shown an overall good performance over the Mel-Frequency Cepstral Coefficients (MFCC) under voice disguise scenario [22].

The present study mainly focusses on studying the effects of non-electronic voice disguises on the performance of the speaker identification system. Here, for training the speaker models, not only normal speaking but voice disguised speech, fast and synchronous speaking were also utilized. For testing, voice disguises, fast, synchronous and repetitive synchronous were used along with the normal speaking speech data. The use of different types of voice disguises for training and testing created a good mismatch, both in speaking style, and channel. It made the speaker identification more challenging. This kind of situation is faced in the real world by the person who is investigating a crime case and dealing with the suspects. Finally, a method which combines: 1) multiple frame rate for obtaining the frames and thereby features, and 2) reliable frame selection at the decision level is proposed, in order to increase the identification accuracy of the system under voice disguise, the ultimate goal of the system. For simplicity, non-electronic voice disguise will be referred to as voice disguise only in the rest of the paper.

This paper is organized as follows. The next section of the paper first presents the frame selection method at the decision level, and then the proposed speaker identification system combining features obtained using multiple frame rate and frame selection at the decision. Section 3 describes the modified/voice disguised speech used in this study. Section 4 describes the database, the experimental setup and then discusses the results, and finally Section 5, concludes the paper with some future insights.

2 Multiple Frame Rates for Feature Extraction and Reliable Frame Selection at the Decision

In the testing, the speech utterance from an unknown speaker is given, and the objective is to find out the true speaker of the utterance from a group of N speakers $\{1, 2, 3, \dots, N\}$, represented by the speaker models $\{\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_N\}$.

In the conventional method, the true speaker amongst the N speakers is one, whose model gives the maximum posterior probability for the feature vector sequence $X = \{x_1, x_2, x_3, \dots, x_T\}$ of the given speech utterance. It is given by the following decision rule [23]:

$$\hat{N} = \arg \max_{1 \leq n \leq N} \sum_{t=1}^T \log p(\mathbf{x}_t | \lambda_n) \quad (1)$$

where, $\mathbf{x}_t$ is a D-dimensional feature vector with T as the total number of feature vectors obtained, $p(\mathbf{x}_t|\lambda_n)$ is the Gaussian mixture density of speaker N, which is represented by the linear weighted sum of M component unimodal Gaussian densities:

$$p(\mathbf{x}_t|\lambda_n) = \sum_{i=1}^M w_i b_i(\mathbf{x}_t) \quad (2)$$

Here, w_i are the mixture weights with $\sum_{i=1}^M w_i = 1$ and $b_i(\mathbf{x}_t)$ are the Gaussian component densities given by:

$$b_i(\mathbf{x}_t) = \frac{1}{(2\pi)^{\frac{D}{2}} \sqrt{\det C_i}} \times \exp\left\{-\frac{1}{2}(\mathbf{x}_t - \mu_i)^T C_i (\mathbf{x}_t - \mu_i)\right\} \quad (3)$$

Here, μ_i represents mean vector and C_i represents covariance matrix of the Gaussian component.

In the proposed method, the true speaker of the given speech utterance is calculated differently and is explained in the following steps [24]:

1. The true speaker is found out on a frame by frame basis instead of the whole speech utterance and is given by:

$$\hat{N} = \arg \max_{1 \leq n \leq N} p(\mathbf{x}_t|\lambda_n) \quad (4)$$

where $p(\mathbf{x}_t|\lambda_n)$ is given by Equation 2.

2. To calculate the reliability or degree of confidence in the decision, the distance between the probability score of the identified speaker and the remaining speakers were found and is given by:

$$D_k = p(\mathbf{x}_t|\lambda_{\hat{N}}) - p(\mathbf{x}_t|\lambda_k) \quad k = \{1, 2, 3, \dots, N\} - \{\hat{N}\} \quad (5)$$

The larger the distance D_k , the more is the confidence in the decision.

3. In this step, the distances D_k are compared with a threshold θ . If $D_k > \theta$, the frame $\mathbf{x}_t$ is kept, otherwise discarded.
4. After step 3, the less reliable frames from the decision point of view are discarded. The remaining frames of the speech utterance will participate in the decision about the true speaker using the conventional decision rule given by Equation 1.

The non-stationary speech utterance is cut into short segments called frames for feature extraction, in which, it exhibits stationary behaviour. The

conventional method typically uses, a frame size of 25ms–30ms with a fixed frame shift of 10ms [22, 23]. But our previous study on voice disguise [25] indicated that changing the frame rate can improve identification accuracy. Therefore, in the proposed method, speaker models were made utilizing a frame size of 25ms and a frame shift of 3ms. For testing, two sets of frames were developed. The first one used frame size of 25ms with frame shift of 10ms and the second set used frame size of 25ms with a frame shift of 3ms. Reliable frames are then selected from these two sets using the method described above. At the end, these two sets are simply combined to make the final frame set for feature extraction and decision making using Equation 1.

3 Modified or Disguised Speech

CHAINS corpus is used to evaluate the different speaker identification experiments conducted for this study [26]. It consists of speech recordings from 36 speakers. The speakers mostly belonged to the same dialect, which raised the difficulty in identification. All the speech recordings were carried out in two sessions separated by 2 month's time. The first session is carried out in a professional studio in a sound proof booth. It used Neumann U87 condenser microphone and the second session is carried out in a quiet office environment using AKG C420 headset condenser microphone.

The speakers along with their normal speech provided 5 modified speech/recordings across the two sessions. The full detail of which can be found in [26]. For the present study, the following three modified/disguised speech along with the normal speech are used:

1. Normal Speech (Norm): In this, the speakers read the given fables and sentences in their normal day to day speaking style. It belonged to the first recording session.
2. Synchronous Speech (Sync): In this, the speakers read the given fables and sentences in synchrony with a co-speaker. Here, both the speaker and the co-speaker modified their voice to match the speaking speed of the other. It produced speech at a relatively slower rate than the Norm speaking. It also belonged to the first recording session.
3. Fast Speech (Fast): In this, the speakers read the fables and sentences at a much faster rate than the Norm speaking. An example of the fast reading has been provided to the speakers, and they tried to match the same speaking speed. It belonged to the second recording session.

Figure 2 Spectrogram of the Norm, Fast and Sync speaking for the speech utterance “If it doesnt matter who wins, why do we keep scores?”[27].

4. Repetitive Synchronous Speech (Rsi): In this, a target speaker’s short speech recording was played in a continuous loop. The speakers joined the played recording and tried to mimick the target. They can hear their own speech and the played target speech, it helped in producing a more closer match with the target. It also belonged to the second recording session.

Out of the three modified/disguised speech, two i.e Sync and Rsi belonged to the imitative category and one i.e Fast belonged to the non-imitative category. The spectrogram of the Norm, Sync and Fast speaking for an utterance are shown in Figure 2. From the figure, it can be seen that, Norm and Sync did not show marked differences but Fast speaking shows differences in comparision with the Norm speaking.

4 Experimental Setup and Results

In this Section, the database, the various speaker identification experiments and the results were discussed.

4.1 Database

Speaker models were trained using Norm, Sync and Fast speech data and tested with unseen data from the Norm, Sync, Fast and Rsi speech of the CHAINS corpus. For training approx. 70sec speech data per speaker were utilized, and for testing 3 utterances of approx. 10sec i.e. making a total of 30sec per speaker were used for Norm, Sync and Fast speech. For Rsi, 4 utterances of approx. 10sec per speaker were used for testing.

Norm and Sync belonged to one recording session and Fast and Rsi belonged to the other recording session.

4.2 Speaker Identification Experiments

Speaker models were developed using 64 component Gaussian mixture models. K-mean clustering algorithm is used to estimate mixture centres and variances. Gaussian models and the weights are adjusted using Expectation Maximization (EM) algorithm [18]. 12 liftered MFCC excluding the 0th coefficient were extracted as features from the speech frames. Liftering is used to rescale the lower and higher order cepstral so that it has similar magnitudes. Cepstral mean removal is utilized for channel compensation.

The following speaker identification experiments were conducted to understand the effects of voice disguise on the speaker identification accuracy:

1. Baseline system utilized frame size of 25ms and frame shift of 10ms. It is referred to as Bsln1. Use of this frame size and frame shift has shown best results, and many research studies with different datasets have utilized these values and is considered a standard in speaker identification [19, 22, 23].
2. Our previous study in which speaker model training utilized only Norm speech has shown that varying the frame rate by changing the frame shift can improve the speaker identification accuracy under voice disguise [20]. A frame size of 25ms and a frame shift of 3ms has shown the best results. Therefore, in this study too, speaker identification experiments were conducted utilizing the frame size of 25ms with a frame shift of 3ms. Here, speaker models utilized training data from three speaking styles, namely, Norm, Sync and fast. It is referred to as Bsln2.
3. Speaker identification experiments were conducted for the proposed system explained in Section 2. It is referred to as Proposed.

4.3 Results and Discussions

The identification accuracy given by:

$$\frac{\text{no. of correctly identified speech utterances}}{\text{total no. of speech utterances tested}} \times 100\% \quad (6)$$

for the speaker identification systems Bsln1 and Bsln2, trained using different speaking style's speech data and tested across unseen different speaking style's speech is tabulated in Table 1. From the Table 1, it can be seen that, when train and test speech data are mismatched both in style and channel (The last two lines of the Table, out of the four tested speech for models trained using a particular speaking style's speech), the identification accuracy decreased markedly. This suggests that speaker identification systems are quite vulnerable to voice disguise. Concentrating on the identification accuracies obtained for the different test speech data, for speaker models trained using a particular speech data. It can be observed that, speaker models trained using Norm speech performed better for the Bsln2 than the Bsln1. Bsln2 showed a relative improvement of 5.35% for the Rsi test speech and a relative improvement of 2.04% for the Fast speech over the Bsln1. On an average across the different test speech data, Bsln2 showed a relative improvement of 1.63% over the Bsln1. For speaker models trained using Sync speech, similar results like the Norm are not found. Here, on an average Bsln1 performed

Table 1 Identification accuracies (%) for Bsln1 & Bsln2 for speaker models trained using different speaking style's speech data across different voice disguises and normal test speech

Train Speech	Test Speech	Bsln1	Bsln2
Norm	Norm	100	100
	Sync	100	100
	Fast	90.74	92.59
	Rsi	77.78	81.94
<i>Average</i>		92.13	93.63
Sync	Norm	96.30	95.37
	Sync	100	100
	Fast	76.85	78.70
	Rsi	72.22	68.37
<i>Average</i>		86.34	85.61
Fast	Fast	99.07	100
	Rsi	95.83	93.75
	Norm	66.67	67.59
	Sync	60.19	62.96
<i>Average</i>		80.44	81.08

better than the Bsln2. For speaker model trained using Fast speech, Bsln2 performed better than the Bsln1 on an average across the different test speech data. These results indicate that, employing different frame rates for feature extraction might improve the identification accuracy under voice disguise.

For the reliable frame selection at the decision level, used in the Proposed system, the calculation of the threshold value θ is required. The value of θ should be calculated after examining all the distances D_k (refer Section 2) of the frames for a given test speech utterance. A larger value of θ compared to the minimum of all the distances D_k means, rejection of more number of frames. These rejected frames will not be able to participate in the decision making process, which might leads to a loss of some important information and thereby decreasing the identification accuracy. So, selection of an optimum value of θ is needed and is tough to determine. Moreover, the value of θ should suitably change with each speech utterance tested. But for this initial study, we have selected a fixed θ value which was decided by examining the test speech utterances from only Rsi speaking style and speaker models trained using Norm speech data. For future studies, we will work on developing an algorithm, where θ value changes according to different test speech utterances and speaking style.

For finding the θ value of this study, we chose different θ values after observing the minimum of all the distances D_k from the various test speech utterances and the identification accuracy obtained for each is plotted. It is shown in the Figure 3. For more clear figure, the log of θ is plotted instead of

Figure 3 Identification accuracies (%) obtained for the different threshold value θ for Rsi speech test data.

Table 2 Identification accuracies (%) for different systems across different voice disguises and normal test speech

Test Speech	Bsln1	Bsln2	Proposed
Norm	100	100	100
Sync	100	100	100
Fast	90.74	92.59	92.59
Rsi	77.78	81.94	83.33
<i>Average</i>	92.13	93.63	93.98

θ . The θ value which gave the highest identification accuracy, i.e. $1.411e-42$, is decided as the final θ value for the rest of the experiments.

Table 2 shows the identification accuracies obtained with speaker models trained using Norm speech and tested across different speaking style/disguised speech along with Norm speech for Bsln1, Bsln2 and Proposed system. It is clear from Table 2 that, on an average across different test speech data, the Proposed system outperformed both the Bsln1 and Bsln2. The Proposed system performed quite good for the Rsi speech test, it attained a relative improvement of 4.35% and 1.70% over the Bsln1 and Bsln2, respectively.

5 Conclusions

This paper studies the effect of three different types of voice disguises from the CHAINS corpus on the performance of the speaker identification system. Out of the three voice disguises, two belonged to the imitative type, namely, Synchronous and Repetitive Synchronous and one belonged to the non-imitative Fast type. It has been found that the speaker identification system's performance markedly decreases on the presence of voice disguise. Making it worse, the types of options available to a person for voice disguise is huge and not fully explored as it will depend on the individual's unique skills for voice disguise. A speaker identification system employing two different frame rates for feature extraction and reliable frame selection at the decision level has been proposed. On an average, it performed better for the speaker models trained utilizing normal speech and tested across different voice disguises over the baseline methods. It particularly performed good for the Repetitive Synchronous Imitation type of voice disguise. Future studies will focus on developing better algorithms for frame selection at the decision level for voice disguised speech and the involvement of more types of voice disguises.

References

- [1] G. R. Doddington. Speaker recognition – Identifying people by their voices. *Proceedings of the IEEE*, 73(11):1651–1664, 1985.
- [2] A. K. Jain, A. Ross and S. Prabhakar. An Introduction to Biometric Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20, 2004.
- [3] T. Kinnunen, H. Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communications*, 52(1):12–40, 2010.
- [4] J. Campbell, P. Joseph. Speaker recognition: a tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, 1997.
- [5] R. J. Mammone, X. Zhang and R. P. Ramachandran. Robust speaker recognition: a feature-based approach. *IEEE Signal Processing Magazine*, 13(5):58–71, 1996.
- [6] L. ZãŁo and R. Coelho Colored Noise Based Multicondition Training Technique for Robust Speaker Identification. *IEEE Signal Processing Letters*, 18(11):675–678, 2011.
- [7] H. Huang, S. Yao, R. Zhou, Y. Yan. Robust speaker recognition using library of cross-domain variation compensation transforms. *Electronics Letters*, 52(4):321–323, 2016.
- [8] A. Kanagasundaram, D. Dean, S. Sridharan. Improving outdomain PLDA speaker verification using unsupervised inter-dataset variability compensation approach. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015.
- [9] I. Shahn. Speaker identification investigation and analysis in two distinct emotional talking environments In *11th International Conference on Signal Processing* Beijing, China, 2012.
- [10] R. Rodman and M. Powell. Computer recognition of speakers who disguise their voice. In *Proceedings of the International Conference on Signal Processing Applications and Technology (ICSPAT)*, Dallas, Texas, USA, 2000.
- [11] AV Voice Changer Software Diamond 7.0 of AVSOFT CORP. [Online]. Available: <http://www.audio4fun.com/voice-over.htm>.
- [12] Y. W. Lau, D. Tran and M. Wagner. Testing voice mimicry with the YOHO speaker verification corpus. *Knowledge-Based Intelligent Information and Engineering Systems*, 15–21, 2005.
- [13] H. Masthoff. A report on voice disguise experiment. *International Journal of Speech Language and the Law*, 3(1):160–167, 1996.
- [14] S. Ganapathy, S. H. Mallidi, H. Hermansky. Robust feature extraction using modulation filtering of autoregressive models. *IEEE Transactions on Audio, Speech and Language Processing*, 22(8):1285–1295, 2014.

- [15] D. You, B. Qiao, J. Li. The optimized dictionary based robust speaker recognition. *Journal of Signal Processing Systems*, 1–9, 2016, in Press.
- [16] N. Dehak, P. J. kenny, R. Dehak, P. Dumouchel, P. Oullet. Front-end factor analysis for speaker verification. *IEEE Audio, Speech, Language Processing*, 19(4): 788–798, 2011.
- [17] W. Endres, W. Bambach and G. Flosser. Voice spectrograms as a function of age, voice disguise and voice imitation. *The Journal of the Acoustical Society of America*, 49(6):1842–1848, 1971.
- [18] A. R. Reich, K. L. Moll and J. F. Curtis. Effects of selected voice disguises upon spectrographic speaker identification. *The Journal of the Acoustical Society of America*, 60(4):919–925, 1976.
- [19] C. Zhang, T. Tan. Voice disguise and automatic speaker recognition. *Forensic Science International*, Elsevier, 175(2):118–122, 2008.
- [20] M. Farrus, M. Wagner, J. Anguita and J. Hernando. Robustness of prosodic feature to voice imitation. In *INTERSPEECH*, Brisbane, Australia, 2008.
- [21] A. Leemann, M. -J. Kolly. Speaker-invariant suprasegmental temporal features in normal and disguised speech. *Speech Communications*, 75:97–122, 2015.
- [22] M. Grimaldi and F. Cummins. Speech style and speaker recognition: a case study. In *INTERSPEECH*, Brighton, U.K, 2009.
- [23] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, 1995.
- [24] S. Prasad and R. Prasad. Reliable frame selection for speaker identification under voice disguise scenario. In *Wireless VITAE*, Hyderabad, India, 2015, in Press.
- [25] S. Prasad, Z. - H. Tan and R. Prasad. Multi-frame rate based multiple-model training for robust speaker identification of disguised voice. In *16th International Wireless Personal Multimedia Communications (WPMC)*, New Jersey, USA, 2013.
- [26] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko. The CHAINS Corpus: CHARacterizing INDividual Speakers. In *Proceedings of SPECOM*, St. Petersburg, Russia, 2006.
- [27] S. Prasad, Z. -H. Tan, R. Prasad. Multistyle training and fusion for speaker identification of disguised voice. In *1st International Conference on Communications, Connectivity, Convergence, Content and Cooperation (IC5)*, Mumbai, India, Dec. 2013.

Biographies

S. Prasad received her M.E degree in Electrical & Electronics from Birla Institute of Technology, Mesra, Ranchi, India in 2007, and is working as Asst. Professor at the Department of Electronics & Communications since then. She is currently pursuing her Ph.D. degree in the Center for TeleInfrastruktur (CTIF) Section, Department of Electronics Systems, Aalborg University. She also holds a B-level certificate in the National Mathematics Olympiad Contest (1995) and is a recipient of the Erasmus Mundus scholarships for pursuing her Ph.D. degree.

Her research interest lies in the area of Speaker Identification, Brain Computer Interface (BCI) and Digital Electronics. She has also reviewed research papers for Computer Speech and Language, and Wireless Personal Communications Journal.

Z.-H. Tan received the B.Sc. and M.Sc. degrees in electrical engineering from Hunan University, Changsha, China, in 1990 and 1996, respectively, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1999.

He has been an Associate Professor with the Department of Electronic Systems, Aalborg University, Aalborg, Denmark, since 2001. He was a Visiting Scientist with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA; an Associate Professor with the Department of Electronic Engineering, Shanghai Jiao Tong

University; and a Post-Doctoral Fellow with the Department of Computer Science, Korea Advanced Institute of Science and Technology, Daejeon, Korea. His current research interests include speech and speaker recognition, noise-robust speech processing, multimedia signal and information processing, human robot interaction, and machine learning.

R. Prasad is currently the Director of the Center for Teleinfrastruktur (CTIF), Aalborg University, Aalborg, Denmark, and holds the chair of wireless information and multimedia communications. He is the Founding Chairman of the Global ICT Standardisation Forum for India, established in 2009. He has published over 700 technical papers, contributed to several books and has authored, coauthored, and edited over 20 books.