Speaker Recognition Robustness to Voice Conversion

Mireia Farrús, Daniel Erro, and Javier Hernando

TALP Research Centre, Department of Signal Theory and Communications
Universitat Politècnica de Catalunya, Barcelona
{mfarrus,derro,javier}@gps.tsc.upc.edu

Abstract. Security systems relying on voice identification can be threatened by human voice imitation or synthetic voices. As voice conversion can be seen as a sort of voice imitation, this paper analyses the performance of an automatic speaker identification system by using converted voices in order to know how vulnerable such systems are to this kind of disguise. The experiments are conducted by using intra-gender and cross-gender conversions between two males and two females. The results show that, in general terms, the system is more robust to intra-gender converted voices than to cross-gender ones.

Key words: speaker identification, voice conversion, robustness

1 Introduction

Voice imitation and other types of disguise are potential threats to security systems that use automatic speaker recognition; therefore, several studies have been performed in order to test the vulnerability of speaker recognition systems against imitation by human or synthetic voices.

Automatic voice conversion is the modification of a speaker voice —called *source speaker*— in order to make it being perceived as if another speaker —*target speaker*— had uttered it. Given thus two speakers, the aim of a voice conversion system is to determine a transformation function that *converts* the speech of the source speaker (from which usually a complete database is available) into the speech of the target speaker (from which normally few data are available), replacing the physical characteristics of the voice without altering the message contained in the speech [1, 2].

Several studies have been done to test the vulnerability of speaker recognition systems against voice disguise and imitations by human or synthetic voices. An experiment reported in [3] tried to deceive a state-of-the-art speaker verification system by using different types of artificial voices created with client speech. Other works related to the vulnerability of automatic recognition systems to specifically created synthetic voices can be found in [4] and [5], where the impostor acceptance rate is increased by modifying the voice of an impostor in order to target a specific speaker.

This paper analyses the robustness of an automatic speaker recognition system against converted voices. The conversion system used to get such converted voices comes up from the improvement of a synthesis system based on the harmonic plus stochastic model [6], which uses frames of fixed length, and where a conversion module has been implemented. The performance of the systems has been demonstrated to be notable, even when no training parallel corpus is available. This is partly due to the fact that the system takes advantage of the high flexibility of the harmonic plus stochastic model in order to minimise the errors derived of the signal reconstruction from their already modified parameters [6].

Next, the voice synthesis system and the voice conversion method are introduced, and the voice conversion database used in the experiments is described in section 3. In order to analyse the robustness of an automatic speaker recognition system against converted voices, the system is tested against both original and converted voices (section 4), so that the comparison will allow to see if the performance gets worse by using voice conversion. Finally, conclusions are presented in section 5.

2 Description of the voice conversion system

The aim of voice conversion systems is to modify the voice produced by a source speaker, for it to be perceived by listeners as if it had been uttered by a target speaker. During the training phase, given a speech database recorded from specific source and target speakers, the system has to determine the optimal transformation for converting one voice into the other one. First, the involved speech signals are frame-by-frame analysed, according to a certain speech model that allows signal manipulation. Then, each analysed frame is translated into a fixed number of parameters with good conversion properties. Finally, after finding the correspondence between the acoustic characteristics of the speakers, the transformation function is learnt. During the conversion phase, the system applies such function to convert new input utterances of the source speaker. Fig. 1 shows the general architecture of a voice conversion system.

The speech model chosen for analysis, transformation and reconstruction of signals is the harmonic plus stochastic model (HSM) [6], which provides high quality speech reconstruction and allows the manipulation of both waveform and spectrum in a very flexible way. Moreover, the model is compatible with many voice transformation methods. The harmonic component captures the part of the signal that is similar to a periodic waveform, and it is characterized by the frequencies, the amplitudes and the phases of the harmonically related sinusoids, whereas the stochastic component containing all the non-sinusoidal signal components is modelled by means of LPC filters. During the analysis, all these features are measured at a constant frame rate. During synthesis, the frames are reconstructed and overlapped.

Converting voices directly from the HSM parameters (amplitudes, frequencies, phases and stochastic LPC filters) is extremely complicated. Instead, the problem can be decomposed into three different sub-problems: pitch conversion,

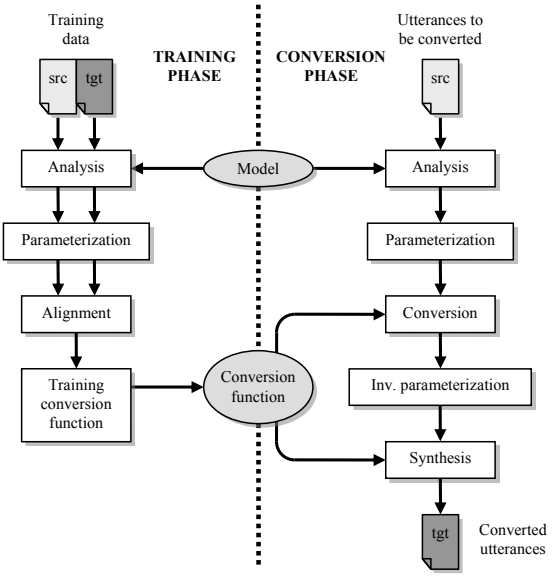

Fig. 1. General architecture of a voice conversion system.

harmonic conversion and stochastic conversion. Since both pitch and stochastic component are represented by very simple parameters (a scalar and an all-pole filter, respectively), the parameterisation task is narrowed to translate the harmonic component into an all-pole filter [7]. Before applying spectral conversion techniques, the harmonic and stochastic all-pole filters are transformed into their associated line spectral frequencies (LSFs) [8], which have very good properties for linear transformations. In order to reconstruct the speech signal from converted LSF vectors, they need to be transformed back into all-pole filters. The stochastic part does not need any extra processing, whereas the harmonic all-pole filter has to be sampled in the frequency domain at multiples of the converted pitch, so that new amplitudes and phases are obtained.

If a parallel training corpus (where the same sentences are uttered by both source and target speakers) is available, the alignment process is simplified and the accuracy of the voice conversion system is increased. In order to train adequate voice conversion functions, a correspondence must be established between the parameter vectors representing the speech frames of the source speaker and those of the target speaker. The method chosen for alignment of source and target frames gives very good results despite its simplicity [9] and consists of the following steps: (i) the boundaries of the phonemes are determined by automatic segmentation based on hidden Markov models, (ii) the phoneme boundaries are used as anchor points to establish a piecewise linear time-warping function for the source-target pairs of parallel sentences, and (iii) each acoustic source vector is paired with the closest target neighbour in the warped time scale.

The method used for spectral envelope conversion is a particular implementation of the GMM-based solution proposed by Stylianou [10] and improved by Kain [11]. It is known that the transformation of the voiced sounds (in which the harmonic component exists) is much more important for voice conversion than the transformation of the unvoiced sounds; therefore, only the voiced frames are transformed, so that only the aligned frame pairs where both members are voiced are considered for training. The spectral conversion method used in this paper consists in applying a GMM-based transformation function to the harmonic LSF vector, and then predicting the stochastic LSF vector from the transformed harmonic one only at voiced frames.

After the alignment and during the training phase, the acoustic mapping between the source speaker and the target speaker is given by a set of frame pairs of the form $\{x_h, x_s\} \leftrightarrow \{y_h, y_s\}$, where the sub-index h denotes the LSF vector of the harmonic component and s denotes the LSF vector of the stochastic component. From now on, and for simplicity, x_h and y_h will be called simply x and y. The paired p-dimensional LSF vectors x and y are concatenated together to form $2p$-dimensional vectors $z = [x^T y^T]^T$. Then, a GMM given by the weights $\{\alpha_i\}$, mean vectors $\{\mu_j\}$ and covariance matrices $\{\Sigma_i\}$ of m different Gaussian components is estimated from the set of vectors $\{z\}$ by means of the expectation-maximization algorithm. Given the relationship between vectors:

$$\mu_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix}, \qquad \Sigma_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix}, \tag{1}$$

the probability of a vector x belonging to the ith Gaussian component of the model $p_i(x)$ can be expressed as:

$$p_i(x) = \frac{\alpha_i N(x, \mu_i^x, \Sigma_i^{xx})}{\sum_{j=1}^{m} \alpha_j N(x, \mu_j^x, \Sigma_j^{xx})} \tag{2}$$

where $N(.)$ denotes the Gaussian distribution. Now, the following transformation function can be applied:

$$F(x) = \sum_{i=1}^{m} p_i(x) \left[\mu_i^y + \Sigma_i^{yx} (\Sigma_i^{xx})^{-1} (x - \mu_i^x) \right]. \tag{3}$$

Under the assumption that the stochastic component is highly correlated with the harmonic component in voiced frames, a stochastic envelope prediction function can be learnt using the training speech frames of the target speaker. Once the transformation function for the harmonic component is trained, all the harmonic-stochastic vector pairs of the form $\{y, y_s\}$ and the target speaker's acoustic model given by $\{\alpha_i, \mu_i^y, \Sigma_i^{yy}\}$ can be used for calculating the m vectors $\{\nu_i\}$ and matrices $\{\Gamma_i\}$ that minimise the error of the following prediction function:

$$y_s = \sum_{i=1}^{m} p_i^y(y) [\nu_i + \Gamma_i (\Sigma_i^{yy})^{-1} (y - \mu_i^y)] \tag{4}$$

During the conversion phase, the prediction function is applied to the converted harmonic LSF vector $F(x)$ instead of y.

With regard to pitch level conversion, a basic adaptation between speakers gives good enough results in most of the cases, especially when the speech signals used for test are emotionally neutral. Since $\log(f_0)$ is well represented by a normal distribution, the pitch level is well converted by applying the following transformation based on replacing the mean and variance of the distribution:

$$\log f_0' = \mu^y_{\log f_0} + \frac{\sigma^y_{\log f_0}}{\sigma^x_{\log f_0}} \Big(\log f_0 - \mu^x_{\log f_0} \Big) \; . \tag{5}$$

The full voice conversion system described here is reported to provide very good results in terms of similarity between converted and target voices, although the quality of the converted signals is affected by a certain over-smoothing effect caused by the statistical transformation procedure [2].

3 Voice conversion database

The database used for voice conversion was made available by UPC for the evaluation campaigns of the TC-STAR project [12]. The voice conversion corpora contain around 200 sentences in Spanish and 170 in English —although only the Spanish ones were used in these experiments— uttered by four different professional bilingual speakers, 2 males and 2 females. The average duration of the sentences is 4 seconds, so that about 10-15 minutes of audio were available for each speaker and language. The sentences uttered by the speakers are exactly the same, so that parallel training corpora can be used for training voice conversion functions, In addition, the sentences were recorded as mimic sentences. This means that there were no significant prosodic differences between speakers, since they all were asked to imitate the same prerecorded pattern with neutral speaking style for each of the sentences.

4 Identification Experiments

First of all, the original data set consisting of all four voices described in the previous section was divided in three sets of sentences. The first set was set aside to train the transformation function of the conversion system, and the second and third set of sentences were used to train and test the automatic recognition system, respectively. Each of the four original voices was converted to the rest of the voices. Since there are 12 pairs of source-target voices, a set of 12 converted voices was obtained: four sets corresponding to intra-gender conversions (female to female and male to male conversions), and eight sets corresponding to cross-gender conversion (female to male and male to female conversions). Each set of converted voices consisted of 100 sentences.

The transformation function for the conversion system was trained using 10, 30 and 80 pairs of source-target sentences. Other 10 original sentences were used

to train each of the four speaker models of the recognition system, and 100 more original sentences, together with the converted sentences, were used for testing. The recognition system utilised in the identification experiments was a conventional 32-component GMM system, using short-term feature vectors consisting of 20 MFCC with a frame size of 24 ms and a shift of 8 ms. The corresponding delta and acceleration coefficients were also included.

In order to test the performance of the recognition system, a preliminary experiment was conducted by using only the original voices. Table 1 shows the corresponding identification matrix, where 100 sentences of each original voice were identified from the closed set of four speaker models. Since it was a rather simple experiment that used a low amount of speakers, it gave a high performance, leading to a percent identification of 100% in three of the four voices. Only one of the males (M1) was confused once with the other male (M2), which suggests —given the high performance of the system— that both male voices are characterised by a significant degree of similarity.

Table 1. Identification matrix for two male (M) and two female (F) original voices.

	F1	**F2**	**M1**	**M2**
F1	100	0	0	0
F2	0	100	0	0
M1	0	0	99	1
M2	0	0	0	100

The identification experiments were conducted by testing both intra-gender and cross-gender converted voices. The system tried to identify 100 sentences of each converted voice again from the closed set of four speaker models. Moreover, three sets of converted voices were identified, according to the sentences used in training the transformation function (10, 30 or 80), in order to see how the amount of training data in the conversion phase influenced the performance of the recognition system.

Table 2. *Source* (a), *target* (b) and *other* (c) identifications using 10 sentences in training the transformation function.

Source voice	**Target voice**			
	F1	**F2**	**M1**	**M2**
F1	-	0	0	0
F2	0	-	0	0
M1	0	0	-	0
M2	0	16	93	-

(a) *Source* identification.

Source voice	**Target voice**			
	F1	**F2**	**M1**	**M2**
F1	-	-	46	100
F2	100	-	98	100
M1	100	98	-	100
M2	100	84	7	-

(b) *Target* identification.

Source voice	**Target voice**			
	F1	**F2**	**M1**	**M2**
F1	-	-	54	0
F2	0	-	2	0
M1	0	2	-	0
M2	0	0	0	-

(c) *Other* identification.

Tables 2, 3 and 4 show the identification results corresponding to the number of sentences used to train the transformation function: 10, 30 and 80, respectively. (The converted F1_to_F2 voices by using 10 training sentences were damaged and not available at the time of doing the experiments). In each table, three types of identification are distinguished: (a) **source**: where the converted voice was identified as its corresponding source speaker, (b) **target**: where the converted voice was identified as its corresponding target speaker, and (c) **other**: where the converted voice was identified as a speaker other than the corresponding source and target speakers.

Table 3. *Source* (a), *target* (b) and *other* (c) identifications using 30 sentences in training the transformation function.

Source voice	Target voice F1	F2	M1	M2
F1	-	-	0	0
F2	0	-	0	0
M1	0	0	-	0
M2	0	9	92	-

(a) *Source* identification.

Source voice	Target voice F1	F2	M1	M2
F1	-	99	43	100
F2	100	-	95	100
M1	100	98	-	100
M2	100	91	8	-

(b) *Target* identification.

Source voice	Target voice F1	F2	M1	M2
F1	-	1	57	0
F2	0	-	5	0
M1	0	2	-	0
M2	0	0	0	-

(c) *Other* identification.

Table 4. *Source* (a), *target* (b) and *other* (c) identifications using 80 sentences in training the transformation function.

Source voice	Target voice F1	F2	M1	M2
F1	-	0	0	0
F2	0	-	0	0
M1	0	0	-	0
M2	0	5	72	-

(a) *Source* identification.

Source voice	Target voice F1	F2	M1	M2
F1	-	100	87	100
F2	100	-	100	100
M1	100	99	-	100
M2	100	95	28	-

(b) *Target* identification.

Source voice	Target voice F1	F2	M1	M2
F1	-	0	13	0
F2	0	-	0	0
M1	0	1	-	0
M2	0	0	0	-

(c) *Other* identification.

The identification results corresponding to 30 training sentences are also plotted in Fig. 2, in which the identification types are also represented by different colours: green, yellow and red for *source*, *target* and *other* identifications, respectively.

Regarding intra-gender identification, the results show that most of the converted voices were identified as their target voices, so that the recognition system failed in identifying the converted voice as the real source voice. Nevertheless, there is one case in which the performance of the system was better —or, in other words, where the voice conversion was not so successful. This is the conversion of the second male to the first male (M2_to_M1). Most of the speakers were identified as the original source voice (M2) instead of as the target voice (M1).

This could probably be explained by the fact that speaker M2 may be highly characterised by his unvoiced segments, and since these are not converted by the system, this unvoiced characteristics still remain in the converted M2_to_M1 voice. However, the identification as the source voice —which will be referred to as *correct identification* by convention— decreases as the amount of conversion training data increases.

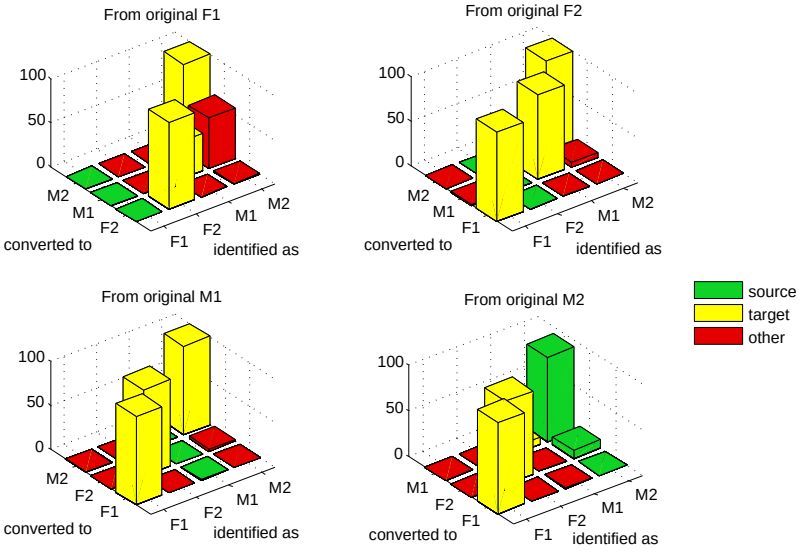

Fig. 2. Identification of each converted voice using 30 sentences in the transformation function. Green, yellow and red bars indicate *source*, *target* and *other* identification, respectively.

It seems thus that the conversion system has difficulties in converting M2 to M1, which could be explained by the fact (seen in Table 1) that both M1 and M2 seem to be similar. However, the reverse phenomenon (M1_to_M2 identified as M1) is not observed in these experiments. Moreover, since the converted F1_to_F2 voice is strangely identified as the male speaker M2 in Table 1, it seems that the recognition system has a slight tendency to identify any speaker as M2.

On the other hand, half of the eight sets of cross-gender converted voices lead to a *miss identification* and *correct conversion* equaling 100%; i.e. not only were the converted speakers not identified as the corresponding source speaker (*miss identification*) but they also were identified as the corresponding target speaker (*correct conversion*).

The other half of the cross-gender conversions were not completely recognised as their corresponding target voices. These are those conversions trying to convert a female speaker to M1 and a male speaker to F2. All the errors are a miss conversion to speaker M2, except in the conversion M2_to_F2, where this

errors can be seen, in fact, as a correct identification of the speaker M2. The worse results are found in the F1_to_M1 conversion, where the tendency of the system to identify speakers as if they were speaker M2 is summed to the hypothetic similarity between M1 and M2 seen in Table 1. In all cases, however, an increase of the correct conversion is observed when the transformation function is trained using 80 sentences.

Summarising, Table 5 shows the types of identification generated by both intra-gender and cross-gender conversions, which are also plotted in Fig. 3. In general terms, intra-gender conversion tends to be identified as its corresponding source speaker in higher degree than cross-gender conversion. On the other hand, cross-gender conversion tends to be more *successful* (speaking in conversion terms) than the intra-gender one, since the percentage of target identification is greater. Nevertheless, cross-gender conversion also leads to a higher percentage of *other* identification; ie. an erroneous conversion in which the converted voice is not identified as either of the source and target speakers.

Table 5. Identification in percent of intra-gender and cross-gender conversions depending on the type of identification generated (*source*, *target* and *other*), where the transformation function has been trained using 30 sentences.

Conversion type	**Source**	**Target**	**Other**
Intra-gender	23.0%	76.7%	0.3%
Cross-gender	1.1%	90.9%	8.0%

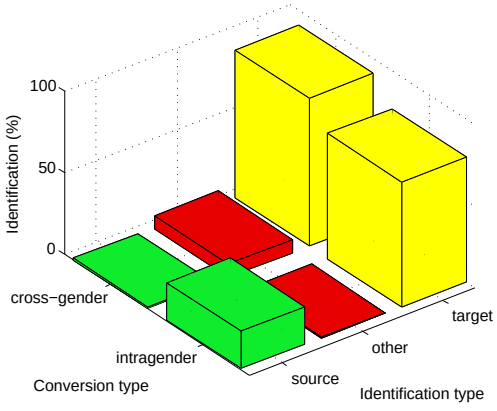

Fig. 3. Identification of intra-gender and cross-gender conversions using 30 training sentences depending on the type of identification generated (*source*, *target* and *other*).

5 Conclusions

In this paper, a set of experiments has been proposed in order to analyse the behaviour of an automatic speaker recognition system against converted voices, using two male and two female voices and several amounts of sentences to train the transformation function. In these experiments, most of the converted voices were identified as their corresponding target speaker; however, they failed sometimes to deceive the system and the source voice was recognised, especially in the intra-gender conversions, which leads to think that the recognition system may be more robust to these kind of conversions than the cross-gender ones. The current results also point out that some voices are more difficult to convert than others, and that the correct identification decreases as the amount of conversion training data increases. Nevertheless, the amount of training data is small enough to interpret the results with extreme caution.

References

1. Duxans, H. Voice Conversion applied to Text-to-Speech systems. PhD Thesis, Universitat Politècnica de Catalunya, Barcelona (2006)
2. Erro, D., Moreno, A.: Sistema de síntesis armónico/estocástico en modo pitch-asíncrono aplicado a conversión de voz. In: Proceedings of the IV Jornadas en Tecnología de Habla. Zaragoza (2006)
3. Lindberg, J., Blomberg, M.: Vulnerability in speaker verification: A study of technical impostor techniques. In: Proceedings of the Eurospeech, pp. 1211–1214. Budapest, Hungary (1999)
4. Masuko, T., Tokuda, K., Tobayashi, T. Imposture using Synthetic Speech Against Speaker Verification Based on Spectrum and Pitch. In: Proceedings of the ICSLP. Beijing, China (2000)
5. Matrouf, D., Bonastre, J.F., Fredouille, C. Effect of speech transformation on impostor acceptance. In: Proceedings of the ICASSP. Tolouse, France (2006)
6. Erro, D., Moreno, A., Bonafonte, A. Flexible Harmonic/Stochastic Speech Synthesis. In: Proceedings of the 6th SSW6. Bonn, Germany (2007)
7. El-Jaroudi,A., Makhoul, J. Discrete All-Pole Modeling. In: IEEE Transactions on Signal Processing (1991)
8. Itakura, F. Line spectrum representation of linear predictive coefficients of speech signals. In: Journal of the Acoustical Society of America, vol. 57 (1975)
9. Duxans, H., Erro, D., Pérez, J., Diego, F., Bonafonte, A., Moreno, A. Voice Conversion of Non-Aligned Data using Unit Selection. In: Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation, Barcelona (2006)
10. Stylianou, Y. Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification. PhD Thesis, École Nationale Supérieure des Télécommunications. Paris, France (1996)
11. Kain, A. High resolution voice transformation. PhD Thesis, OGI School of Science and Engineering (2001)
12. Bonafonte, A., Höge, H., Kiss, I., Moreno, A., Ziegenhain, U., van den Heuvel, H., Hain, H.U., Wang, X.S., Garcia, M.N. TC-STAR: Specifications of language resources and evaluation for speech synthesis. In: Proceedings of the LREC. Genoa, Italy (2006)