The International Journal of Speech, Language and the Law

IJSLL (PRINT) ISSN 1748-8885
IJSLL (ONLINE) ISSN 1748-8893

Article

Automatic speaker recognition as a measurement of voice imitation and conversion

Mireia Farrús, Michael Wagner, Daniel Erro and Javier Hernando

Abstract

Voices can be deliberately disguised by means of human imitation or voice conversion. The question arises to what extent they can be modified by using either method. In the current paper, a set of speaker identification experiments are conducted; first, analysing some prosodic features extracted from voices of professional impersonators attempting to mimic a target voice and, second, using both intragender and cross-gender converted voices in a spectral-based speaker recognition system. The results obtained in the current experiments show that the identification error rate increases when testing with imitated voices, as well as when using converted voices, especially the crossgender conversions.

KEYWORDS IMITATION; VOICE CONVERSION; PROSODY; JITTER; SHIMMER; SPEAKER RECOGNITION

Affiliations

Mireia Farrús, Daniel Erro and Javier Hernando: Universitat Politècnica de Catalunya, Spain

Michael Wagner: University of Canberra, Australia

email: mfarrus@gps.tsc.upc.edu

IJSLL VOL 17.1 2010 119–142

doi : 10.1558/ijsll.v17i1.119

equinox www.equinoxpub.com

equinoxonline

1 Introduction

Human voice imitation can be found in three main aspects of daily human communication: language acquisition, impersonation for entertainment, and voice disguise for concealing a personal identity (Zetterholm, 2003). Nevertheless, human imitation is not the only way to imitate others' voices: automatic voice conversion is the modification of a speaker's voice (called source speaker) in order to create the perception that it was uttered by another speaker (target speaker). Given thus two speakers, the aim of a voice conversion system is to determine a transformation function (TF) that converts the speech of the source speaker into the speech of the target speaker, replacing the speaker characteristics of the voice without altering the message contained in the speech (Duxans, 2006; Erro, Moreno and Bonafonte, 2007). According to Rodman's classification of disguised voices (Rodman, 1998), human imitation intended for concealing identities and automatic converted voices are known as non-electronic and electronic deliberate modifications, respectively.

Several studies have been done in order to test the performance of speaker recognition systems when using voice disguise and imitations by human or synthetic voices. An experiment reported in Lindberg and Blomberg (1999) tried to deceive a state-of-the-art speaker verification system by using different types of artificial voices created from voices of speakers stored in the database. Some works related to the vulnerability of automatic speaker recognition systems to specifically created synthetic voices can be found in Masuko, Tokuda and Tobayashi (2000) and Matrouf, Bonastre and Fredouille (2006), where the impostor acceptance rate is increased by modifying the voice of an impostor in order to target a specific speaker.

Other studies have dealt with the effects of common types of voice disguise against automatic speaker recognition systems. Some preliminary experiments reported in Künzel, González-Rodríguez and Ortega-García (2004) about the effects of increased voice pitch, lowered pitch and pinching the nose while speaking, showed that the performance of an automatic speaker recognition system was degraded by all three modes of disguise, where the lowered pitch mode presented the smallest degradation.

The way how automatic speaker recognition reacts to human imitation has been tested in some recent studies (Lau, Tran and Wagner, 2005; Lau, Wagner and Tran, 2004), where it was shown that an impostor who knows a client speaker of the database with a similar voice to his own voice, could attack the system. Some studies concern themselves with the speaker and dialect imitations research considering both automatic speaker recognition and human speech perception. Both approaches were used, for instance, on research conducted by Sullivan and Pelecanos (2001) and Zetterholm, Blomberg and Elenius (2004). The work by Sullivan and Pelecanos (2001) showed that the recognition

system was capable of classifying the mimic attacks more appropriately than human listeners, whereas the work by Zetterholm et al. (2004) found a minimal correlation between the speaker verification system they used and their human listeners in how they judged the imitations in closeness to the target speaker.

The aim of the current paper is twofold. First, the paper tries to quantify how a human being is able to approximate others' voices, and to what extent some selected prosodic and acoustic features are imitated, by analysing the voices of two professional imitators trying to impersonate several well-known politicians. An objective measurement of the impersonation success is performed by using an automatic speaker recognition system based on prosodic and acoustic features. Second, the paper deals with an experiment that uses imitated voices by means of a specific automatic voice conversion system. As in the previous experiment, an automatic speaker recognition system is used to test, in an objective way, the quality of such converted voices, i.e. how closely they reach their corresponding target voice.

This paper is structured as follows. In the next section, a brief description of the state of the art in human imitation and the prosodic features analysed in our experiments is presented. In Section 3, the voice conversion system used to manipulate a set of source speakers is described. Sections 4 and 5 deal with the experiments related to human imitations and voice conversions, respectively and, finally, conclusions are presented in Section 6.

2 Human voice imitation and prosodic features

It is well known that voice is characterised by a high degree of variability due to several non-deliberate factors such as ageing, intoxication, illness or emotional stress. Moreover, one's voice can be also deliberately modified, such as speaking in falsetto or feigning a speech defect or foreign accent. However, deliberate modifications vary across speakers. A study reported in Künzel (2000), for instance, showed sex-related differences in the strategies employed for disguise by men and women.

As was stated in Section 1, voice imitation can be found in the areas of language acquisition, impersonation for entertainment, and voice disguise for concealing a personal identity (Zetterholm, 2003). Imitation in language acquisition is used mainly for learning both native and foreign languages, but also for accommodation of the speaking manner in a community. Changing one's own dialect or sociolect, for example, to the ones spoken by a community, can be seen as a way of achieving integration into a social group. According to Markham (1997), imitation –in the sense of acquisition– can be manifested in several ways: repetition of words, reproduction of syntactic structures, phonetic reproduction, etc., phonological and phonetic acquisition being the most clearly imitative processes.

When imitation has the aim of reproducing another speaker's voice and speech behaviour, it is usually called impersonation (Markham, 1997). Impersonators are normally found in entertainment environments, and they have the ability to pretend successfully to be someone else, being able to identify, select and imitate characteristic features of the target speaker. For entertaining taking place on stage, the impersonator normally copies the body language and other non-vocal features of the target person as a complement to the vocal imitation. On the other hand, when the impersonator is not seen by the audience, it is more important to focus on imitation of vocal features; but wherever the impersonation takes place, the imitator normally tends to focus on the most prominent features and to exaggerate them (Zetterholm, 2003).

When the aim is to hide one's identity, the voice alteration normally involves changes in vocal tract settings, pitch, voice quality, dialect, accent, prosodic patterns, etc. In this case, the individual may not want to imitate any specific person, but simply try to disguise their own voice. However, there are some physiological differences between speakers that cannot be changed; and when these differences are considerably large, it may be difficult to achieve good imitations of another person's voice (Laver, 1994). An extreme case is found between male and female voices, which show differences concerning fundamental frequency, intensity, shape of the glottal wave, etc. (Pittam, 1994). In a study performed by Lass, Trapp, Baldwin, Scherbick and Wright (1982), some speakers were asked to attempt to speak like the opposite sex, but an auditory analysis revealed the actual sex of the speakers. Also, a professional impersonator interviewed by Zetterholm (2003) declared that he found it much easier to imitate the older voices than the younger ones.

The question arises as to whether it is enough to pick out and copy a number of specific voice features from another person, or whether having a similar voice is more important than the feature selection itself. In this sense, the work found in Zetterholm (2003) concludes that impersonators usually capture several aspects of the target voice, and that an imitation may be successful on the whole even if it fails to imitate some features, provided they are successful in impersonating the most prominent ones.

There is a wide variety of parameters that can be used in imitation. Some of these parameters are more linguistic in nature, such as those related to dialect or sociolect or those related to language style and the selection of lexical items. The importance of dialect disguise in speaker recognition, for example, has been pointed out by researchers such as Shuy (1995). Other imitation parameters are more phonetic in nature. The phonetic imitation parameters can be divided into those that belong to the voice source and those belonging to the vocal tract filter. Early automatic speaker recognition systems tended to use only the filter parameters, and this will also be the method used in

Section 5 (cepstral coefficients). More recently, source parameters have been utilised in state-of-the-art recognition systems (Farrús, Garde, Ejarque, Luque, and Hernando, 2006; Peskin, Navrátil, Abramson, Jones, Klusacek, Reynolds and Xiang, 2003; Reynolds, Andrews, Campbell, Navrátil, Peskin, Adami, Jin, Klusacek, Abramson, Mihaescu, Godfrey, Jones and Xiang, 2003).[1] These source parameters relate mainly to the fundamental frequency and power of the speech sounds, which are also commonly classified as 'prosodic' parameters.

Another set of voice source parameters that have been shown to improve speaker verification systems, and that will be used here are jitter and shimmer (Farrús, Hernando and Ejarque, 2007). Jitter and shimmer are not always considered prosodic parameters, but since they relate to fundamental frequency and power (being microvariations in fundamental frequency and power, respectively), it is justified to classify them as prosodic as well. The term prosodic is also appropriate for a set of duration parameters that will be employed here. Therefore, 'prosodic' will be used as a cover term for all the parameters that will be employed in the voice imitation experiments in Section 4. They are the following twelve parameters, which are analysed in a set of imitated and natural voices. These parameters include the length of words, word segments, means, extrema and ranges of the fundamental frequency, as well as jitter and shimmer measurements, as listed below:

- logarithm of number of frames per word
- length of word-internal voiced segments
- length of word-internal unvoiced segments
- log (mean f0)
- log (max f0)
- log (min f0)
- log (range f0)
- f0 slope
- mean f0 absolute slope
- jitter (absolute)
- shimmer (absolute)
- shimmer (apq3)

All these features are based on the prosodic systems described in recent works of the authors (Farrús et al., 2006; Farrús et al., 2007).

The use of pauses as prosodic features was also analysed in Farrús et al. (2006), where it turned out that they were not relevant in conversations between two speakers, since pause length and rate depended considerably on the speaking rate of the other speaker. Moreover, the sentences analysed in the current paper were too short to take pauses into consideration as relevant features; thus, the pause analysis was not included in the feature set.

3 Voice conversion system

Human voices can be disguised by means of human impersonation, but also by means of voice conversion. In both cases, disguise appears as a relevant and real question for forensic considerations, since the aim is to hide or falsify one's own identity. Some voice transformation techniques are as simple as using a handkerchief over the mouth, while some others are as sophisticated as software manipulation in order to compromise someone else (Perrot, Morel, Razik and Chollet, 2009).

This section deals with the effects of using a voice conversion system over several natural voices. The aim of voice conversion is to modify the voice produced by a source speaker, so that it is perceived by listeners as if it had been uttered by a target speaker.

State-of-the-art voice conversion systems focus only on the transformations of the acoustic voice characteristics: short-time spectrum, mean pitch level, and –only in some cases– prosodic contours, and they consist of two phases: training and conversion.

During the training phase, given a speech database recorded from some specific source and target speakers, the system determines the optimal function to convert the source voice into the target one. The training phase of a generic voice conversion system consists of the following steps:

1) Short-term analysis of the training utterances. A vector containing the parameters of the most relevant voice features is extracted from each frame.

2) Alignment of phonetically equivalent source-target vector pairs. In general, the training process is carried out on a parallel corpus, in which both speakers utter the same sentences, so that the phonetic correspondence between source and target frames is easily found by means of alignment techniques.

3) Learning of the transformation function from the vector pairs.

Then, the system applies this function to convert new input utterances of the source speaker during the conversion phase, which consists of the following steps:

1) Frame-by-frame analysis and parameter extraction.

2) Vector transformation using the trained function.

3) Inverse parameterisation and speech reconstruction.

Among all the different types of spectrum transformation found in the literature, the most popular one – and therefore the one chosen for the experiments presented in the current paper – is the linear transformation based on Gaussian mixture models (GMMs). The GMM divides the vector space of the speakers into *m* classes represented by Gaussian distributions given by their mean vector (**μ**), covariance matrix (Σ), and weighting factor (α). A linear statistically-motivated transformation is defined for each class or Gaussian component. The resulting function is then the combination of contributions from all the classes as follows:

$$F(\mathbf{x}) = \sum_{i=1}^{m} p_i(\mathbf{x}) \left[\boldsymbol{\mu}_i^y + \Sigma_i^{yx} \Sigma_i^{xx\,-1} (\mathbf{x} - \boldsymbol{\mu}_i^x) \right] \quad (1)$$

where **x** is the input source vector, $p_i(\mathbf{x})$ is the probability of $\boldsymbol{x}$ belonging to the *i*th acoustic class, and the super-indices *x* and *y* refer to the source and target speakers, respectively. A detailed explanation of the procedures used to estimate all the parameters of this transformation can be found in Stylianou, Cappé and Moulines (1998) and Kain and Macon (1998).

In the pitch level conversion, the most basic procedure consists of a simple adaptation between speakers: since the pitch logarithm *log(f*0*)* is well represented by a normal distribution, the pitch level is well converted by replacing the mean and variance of the source speaker's *log(f*0*)* distribution by those of the target speaker's distribution, as shown in the following expression:

$$\log f0' = \mu_{\log f0}^y + \frac{\sigma_{\log f0}^y}{\sigma_{\log f0}^x} \left(\log f0 - \mu_{\log f0}^x \right) \quad (2)$$

Although the intonation patterns remain the same after modification (only the mean level and the range are corrected), the results are good enough in most cases, especially when the speech signals used for test are emotionally neutral or have a low degree of prosodic expressiveness.

In order to perform the experiments presented in Section 5, a voice conversion system was implemented according to the following technical specifications. The model chosen for analysis, transformation and reconstruction of the speech signals was based on a harmonic plus stochastic decomposition, in which the periodic part of the signal is represented by a sum of harmonic sinusoids (given by their fundamental frequency, amplitudes and phases) and the aperiodic part is represented as filtered white noise. This model is characterised by a high-quality speech reconstruction, and it is compatible with many voice transformation methods, since it allows the manipulation of both waveform and spectrum in a very flexible way. A more detailed description of the model and

its associated signal manipulation procedures can be found in Erro et al. (2007). During voice conversion, the unvoiced frames were left unmodified –since they are not relevant to identify speakers in this kind of systems– whereas the parameters of the voiced frames were translated into constant-length vectors and were converted by GMM-based linear transformations.

4 Experiments on imitated voices

The current experiments explore the ability of professional mimickers to approximate the prosody of their target voices. The study comprises a set of experiments, in which professional voice imitators mimic the voice characteristics of well-known public figures. In each experiment the prosodic parameters shown in Section 2 are measured and compared between the target speaker's voice (target), the imitator's natural voice (i-natural) and the imitator's modified voice (i-modified).

For each i-natural, i-modified and target voice, a vector consisting of the twelve prosodic parameters described in Section 2 was extracted to perform the identification experiments. For each of those parameters, a baseline speaker identification experiment is conducted to establish the error rate in per cent of a speaker identification system that would try to distinguish between the target speaker and the imitator's natural voice on the basis of the single parameter.

Then, a second experiment is conducted –again for each individual prosodic parameter– to establish the error rate in per cent of a speaker identification system that would try to distinguish between the target speaker and the imitator's modified voice, again on the basis of each prosodic parameter. It is the comparison between the two experiments that reveals, for each of the twelve parameters, how much the professional imitator is able to shift the parameter away from his own voice towards the target speaker's voice. In turn, these comparisons establish the vulnerability of the twelve prosodic parameters against intentional voice mimicking by professionally trained impersonators.

4.1 Material

Two male professional imitators, who will be referred to with their initials (cc and qn) took part in these experiments. They have been working as professional imitators on radio and TV for more than five years. They both are Catalan native speakers and have a Central Catalan dialect.

Five well-known male politicians, who will be referred to also with their initials (JB, JR, JS, PM and XT) were used as target speakers. They were between 45 and 64 years old when the recordings were made. JS, PM and XT are Catalan

native speakers from the same dialectal region as the professional impersonators, while the remaining two (JB and JR) are Spanish native speakers with a Castilian Spanish dialect. All Catalan speakers are Spanish-Catalan bilingual speakers, and since the target speakers spoke the same Catalan dialect as the impersonators or standard Castilian accent, no significant dialectal differences existed between impersonators and targets.

The recordings of the target speakers were taken from public radio interviews, made in local radio station's studios. For each target voice, 20 sentences of about 10–20 seconds length were extracted. The imitations and the natural voices of the impersonators were recorded in their own radio station's studio or in an audio studio at the Department of Signal Theory and Communications at Technical University of Catalonia. The advantage of having recordings from radio interviews and the corresponding imitations made in closed studios without video cameras is that, when the impersonator is not seen by the audience, it is more important to focus on voice similarity, since the listener has no clues other than the voice and speech to identify the target speaker (Zetterholm, 2003).

The impersonators were asked to record both imitated and natural voices with the same text as the recordings of the target speakers. Since a read-text recording may result in a lack of spontaneity, the impersonators had been reading the texts before the recordings in order to copy the target voices as naturally as possible. The professional imitators were asked to impersonate those politicians they are used to impersonate on TV; this allowed successful impersonation, as is routinely accepted by a very big audience. Thus, the impersonator qn imitated the politicians JR, PM and XT, and cc imitated JB and JS. Table 1 shows the imitators and the corresponding target speakers together with the mean fundamental frequency of each speaker. The standard deviation is also shown as a margin error. Both impersonators recorded all the extracted sentences of each target with their natural (i-natural) and modified (i-modified) voices. All the transcriptions were manually word-labelled and aligned.

Table 1: Mean f0 of impersonators and target voices

Imitator	f0 (Hz)	Target	f0 (Hz)
cc	121 ± 37	JB	110 ± 44
		JS	85 ± 54
qn	110 ± 23	JR	81 ± 22
		PM	95 ± 67
		XT	87 ± 27

4.2 Experiments

Both impersonators' voices (i-natural and i-modified voices) were recorded in the same session and in the same recording conditions, while target voices were extracted from previous radio recordings. Due to this mismatch and the small number of speakers used in the experiments, the recognition task with a conventional cepstral-based GMM method was not performed. Therefore, only prosodic parameters were considered, since they seem to be more robust to mismatched recordings (Atal, 1972; Carey, Parris, Lloyd-Thomas, Bennett, 1996).

The parameters were extracted using the Praat software for acoustic analysis (Boersma and Weenink, 1992), performing an acoustic periodicity detection based on a cross-correlation method with a window length of 40/3 ms and a shift of 10/3 ms. The mean over all words was computed for each individual feature.

For every set of 20 different sentences, one speaker model was trained for the i-natural voice and one for the target voice. Either five or ten sentences were used for training the models. The remaining sentences, together with the corresponding i-modified sentences, were used for testing. The system was tested using the k-nearest neighbour classifier (with k=1 and k=3), comparing the Euclidean distances of the test feature vector to the k closest vectors of each set of the trained speaker models.

For each of the twelve parameters, a baseline speaker identification experiment was conducted to establish the error rate of a speaker identification system, which tried to identify the target and i-natural voices from the closest set of two speaker models: the mimicker using his natural voice and the corresponding target speaker, both trained using the same set of sentences. Again for each individual parameter, a second experiment was conducted to establish the error rate of an identification system that tried to identify the target and the i-modified voices from the same closed set of two speaker models: the impersonator speaking with his natural voice and his corresponding target speaker. Thus, in each identification experiment, a total number of 150 tests were performed when the models were trained with 5 sentences (5 targets x 2 speakers x 15 sentences) and 100 tests were performed when the models were trained with ten sentences (5 targets x 2 speakers x 10 sentences).

Finally, the fusion of all the individual features was performed in each experiment at the score level. The scores were normalised with z-score normalisation, which transforms the scores into a distribution with zero mean and unitary variance, and fused with the matcher weighting method, where each individual score is weighted by a factor proportional to the recognition rate (Indovina, Uludag, Snelik, Mink and Jain, 2003).

4.3 Results

The identification error rates (IERs) obtained for both baseline and modified systems are presented in per cent in Table 2. The baseline system is tested with the i-natural and target voices, while the modified system utilises the i-modified and target voices for testing. In the modified system, 'identification error' means that the i-modified voice was identified as the target speaker's voice instead of the imitator's own voice.

The error rates are given for the whole prosodic systems; that is, after fusing all the twelve features involved in the experiments. The table shows the results obtained by using five and ten sentences to train the speaker models. In both cases, the error rates when using k=1 and k=3 in the k-nearest neighbour classification are compared.

Table 2: IER (%) obtained for each prosodic system after fusing all the features

Training sentences	1-NN		3-NN	
	baseline	modified	baseline	modified
5	10.3	19.3	8.7	18.3
10	5.0	22.0	11.0	18.0

The results clearly show that, after fusing all the features, the identification error is always increased when using the modified system instead of the baseline system. The largest difference can be seen when using the 1st nearest neighbour as a classifier and 10 sentences for training.

The identification error rates for each isolated feature are plotted in Figure 1, where the dark line corresponds to the IERs of the baseline system and the light one to the IERs of the modified system. In all cases analysed in Table 2, the results for each individual feature were similar; therefore, only one case (the 1st nearest neighbour and 10 sentences for training) is represented in the figure.

The identification error rates increase for all the individual parameters except for the logarithm of the fundamental frequency range (i.e. the difference between the maximum and minimum values of f0), which remains steady – or even decreases, in this case – in the modified system.

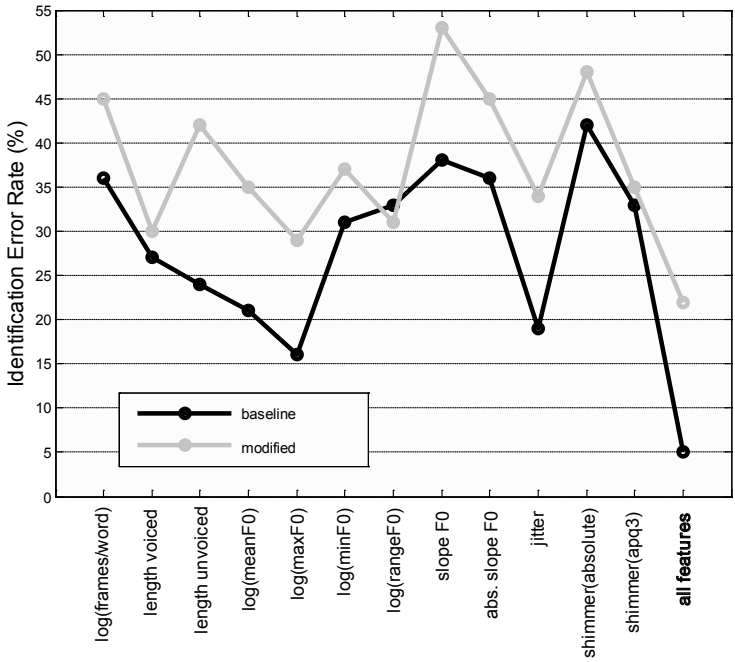

Figure 1: IER (%) for each prosodic feature and fusion using 1st NN and 10 sentences for training

5 Experiments on converted voices

This section analyses the robustness of an automatic speaker recognition system against converted voices. The conversion system derives from the improvement of a synthesis system based on the harmonic plus stochastic model (Erro and Moreno, 2005), 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 (Erro, Moreno and Bonafonte, 2007).

5.1 Material

The database used for voice conversion was made available by the Technical University of Catalonia (UPC) for the evaluation campaigns of the TC-STAR project (Bonafonte, Höge, Kiss, Moreno, Ziegenhain, van den Heuvel, Hain, Wang and Garcia, 2006). The voice conversion corpora contain around 200 sentences in Spanish and 170 in English, uttered by four different professional

bilingual speakers, two males (M1 and M2) and two females (F1 and F2), although only the Spanish sentences were used in the current experiments.

In order to see how similar the selected speakers are, the mean f0 value for each natural voice is shown in Table 3. Moreover, a mean opinion score (MOS) perceptual test was carried out in order to test the similarity between pairs of natural voices. Each listener provided a score in the range 1 to 5, where 1 means that voices where perceived as completely different and 5 means that both voices were perceived as the same voice. Table 3 shows the mean opinion score only for the intragender voice pairs; in the crossgender voice pairs the MOS always equalled 1.

Table 3: Mean f0 of the original voices and mean opinion score of the intragender voice pairs

	f0 (Hz)	voice pair	mean opinion score
F1	207	F1-F2	1.75
F2	214		
M1	111	M1-M2	1.43
M2	123		

The sentences uttered by the speakers are exactly the same, so that parallel training corpora can be used to train 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 pre-recorded pattern with neutral speaking style for each of the sentences.

The average duration of the sentences was four seconds, so that about 10–15 minutes of audio were available for each speaker and language. A detailed description of the corpora, including the recording platform and the speaker selection, can be found in Bonafonte et al. (2006).

5.2 Experiments

First of all, the original data set consisting of all four voices described in Section 5.1 was divided into three sets of sentences. The first set was set aside to train the transformation function of the conversion system, and the second and third sets 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; that is: M1 was converted to M2, F1 and F2; M2 was converted to M1, F1 and F2; F1 was converted to M1, M2 and F2, and F2 was converted to M1, M2 and F1.

Thus, a set of twelve converted voices was obtained: four sets corresponding to intragender conversions (female to female and male to male conversions), and eight sets corresponding to crossgender conversions (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. Ten other 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.

5.3 Results

In order to test the performance of the recognition system, a preliminary experiment was conducted by using only the original voices. Table 4 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 small number of speakers, a high performance was obtained, leading to 100% identification for three of the four voices. Only one of the males (M1) was once confused with the other male (M2), which suggests –given the high performance of the system– that the two male voices are characterised by some degree of similarity.

Table 4: 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 the intragender and crossgender 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 used in the conversion phase influenced the performance of the recognition system.

Tables 5, 6 and 7 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 current experiments). In each table, three types of identification are distinguished:

- *source*: where the converted voice was identified as its corresponding source speaker,
- *target*: where the converted voice was identified as its corresponding target speaker, and
- *other*: where the converted voice was identified as a speaker other than the corresponding source and target speakers.

Table 5: Source (a), target (b) and other (c) identifications using 10 sentences in training the transformation function

	target voice			
source voice	F1	F2	M1	M2
F1	-	-	0	0
F2	0	-	0	0
M1	0	0	-	0
M2	0	16	93	-

(a)

	target voice			
source voice	F1	F2	M1	M2
F1	-	-	46	100
F2	100	-	98	100
M1	100	98	-	100
M2	100	84	7	-

(b)

	target voice			
source voice	F1	F2	M1	M2
F1	-	-	54	0
F2	0	-	2	0
M1	0	2	-	0
M2	0	0	0	-

(c)

Table 6: Source (a), target (b) and other (c) identifications using 30 sentences in training the transformation function

	target voice			
source voice	F1	F2	M1	M2
F1	-	0	0	0
F2	0	-	0	0
M1	0	0	-	0
M2	0	9	92	-

(a)

	target voice			
source voice	F1	F2	M1	M2
F1	-	99	43	100
F2	100	-	95	100
M1	100	98	-	100
M2	100	91	8	-

(b)

	target voice			
source voice	F1	F2	M1	M2
F1	-	1	57	0
F2	0	-	5	0
M1	0	2	-	0
M2	0	0	0	-

(c)

Table 7: Source (a), target (b) and other (c) identifications using 80 sentences in training the transformation function

	target voice			
source voice	F1	F2	M1	M2
F1	-	0	0	0
F2	0	-	0	0
M1	0	0	-	0
M2	0	5	72	-

(a)

	target voice			
source voice	F1	F2	M1	M2
F1	-	100	87	100
F2	100	-	100	100
M1	100	99	-	100
M2	100	95	28	-

(b)

	target voice			
source voice	F1	F2	M1	M2
F1	-	0	13	0
F2	0	-	0	0
M1	0	1	-	-
M2	0	0	0	-

(c)

The identification results corresponding to 30 training sentences are plotted in Figure 2, in which the identification types are represented by different colours: light grey, white and dark grey for source, target and other identifications, respectively.

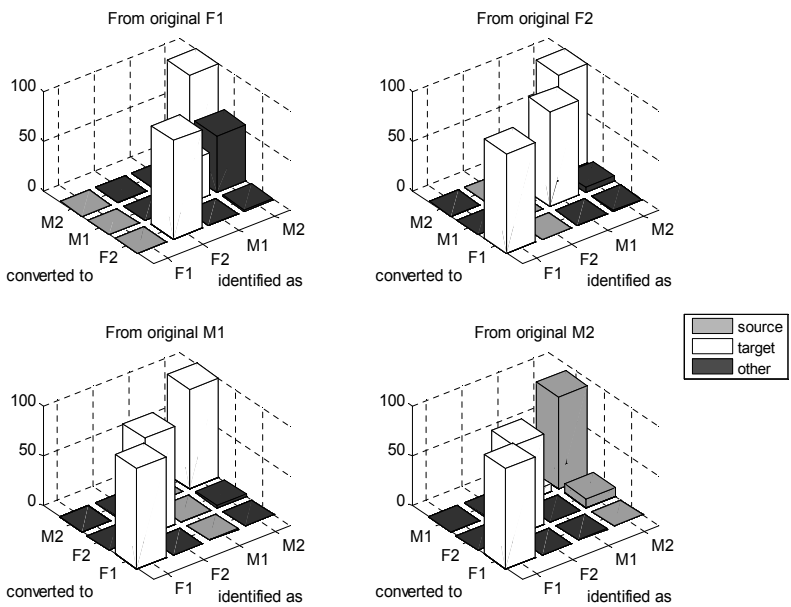

Figure 2: Identification of each converted voice using 30 sentences in the transformation function

Regarding intragender 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): the conversion of the second male to the first (M2_to_M1), where 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.

It seems thus that the conversion system has difficulties in converting M2 to M1, which could be explained by the fact (seen in Table 4) 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, the 'other' speaker in the conversion F1_to_F2 shown in Table 6c turns out to be M2. So 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 crossgender converted voices lead to a 'miss identification' and 'correct conversion' equalling 100%; i.e. not only were the converted speakers not identified as the corresponding source speaker (miss identification) but they were all identified as the corresponding target speaker (correct conversion).

The other half of the crossgender 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 the 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 4. In all cases, however, an increase of the correct conversion is observed when the transformation function is trained using 80 sentences.

Summarising, Table 8 shows the types of identification generated by both intragender and crossgender conversions using 30 training sentences, which are also plotted in Figure 3. In general terms, intragender conversion tends to be identified as its corresponding source speaker in a higher degree than crossgender conversion. On the other hand, crossgender conversion tends to be more 'successful' – speaking in conversion terms – than the intragender conversion, due to the percentage of other identification; i.e. an erroneous

conversion in which the converted voice is not identified as either of the source and target speakers.

Table 8: Percent identification of intragender and crossgender conversions depending on the type of output identification. The TF has been trained using 30 sentences

Conversion type	Source	Target	Other
Intragender	23.0%	76.7%	0.3%
Crossgender	1.1%	90.9%	8.0%

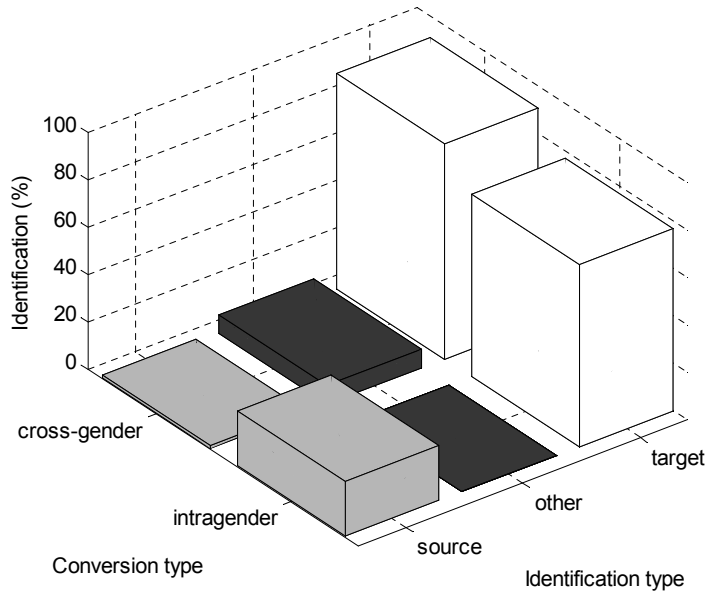

Figure 3: Identification of intragender and crossgender conversions using 30 training sentences depending on the type of output identification

6 Conclusions

A set of experiments was conducted, in which twelve prosodic features were used for speaker identification, and where a professional impersonator attempts to mimic a target voice. For each individual feature, a baseline experiment established models for the target speaker and the natural voice of the impersonator, using a set of training data. A separate set of test data from the target and the impersonator's natural voice was then used to determine the identifica-

tion error rate for the two speakers without attempted impersonation. For each of the twelve features, a second experiment was then conducted, which used the target speaker's test data and the impersonator's modified voice data to determine the identification error rate for the two speakers with attempted impersonation.

For eleven of the twelve features, the identification error rate increased, in some cases greatly, but for the logarithm of the fundamental frequency range the identification error rate remained almost unchanged –in fact, even dropped slightly from 33% to 31%. Fusing the twelve features at the score level resulted in an increase from an identification error rate of 5% for target speakers against the impersonator's natural voice to an identification error rate of 22% for target speakers against the impersonator's modified voice. These results show that the inclusion of prosodic features in the feature set for an automatic speaker recognition system requires careful consideration of the concomitant risk of impersonation, particularly by trained professional imitators.

On the other hand, the behaviour of an automatic speaker recognition system against converted voices has been analysed by using two male and two female voices and different numbers of sentences (10, 30 and 80) 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 intragender conversions. This leads us to think –as expected– that the recognition system may be more robust to intragender conversions than the crossgender ones; thus, it becomes more difficult to deceive the system when converting a voice to another voice of the same sex. 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.

The experiments presented in the current paper show two scenarios where a voice transformation can successfully falsify a personal identity. The features analysed in both experiments are different, according to the idiosyncrasy of the voice transformation used, and the results show which prosodic features are easier to imitate and in which situations voice conversions are more effective. The experiments tried to set a preliminary stage of feature analysis in order to see how easy it can be to forge voices in different forensic situations. Nevertheless, the databases used in both imitation and conversion experiments are small enough to warrant cautious interpretation of the results and further analysis and experimentation would be needed in future work before any more conclusions can be drawn.

About the authors

Mireia Farrús has degrees in Physics and Linguistics from the University of Barcelona. She received her PhD in 2008 from the Universitat Politècnica de Catalunya, in the Department of Signal Theory and Communications, and is now a researcher at the Universitat Oberta de Catalunya (UOC). Her main interests are biometrics, speaker recognition and machine translation.

Michael Wagner received his Diplomphysiker degree from Ludwig-Maximilians-Universität in Munich in 1973, and his PhD from the Australian National University in 1979. Since 1996 he has held the Chair in Computing of the University of Canberra, where he is currently Head of the Discipline of Software Engineering and Director of the National Centre for Biometric Studies. He is the author of more than 140 refereed publications in the field of speech science and technology.

Daniel Erro received the Telecommunication Engineering degree from the Public University of Navarra in 2003, and his PhD degree from the Technical University of Catalonia, Barcelona, in 2008. He is now with the University of the Basque Country. His main research interests include speech analysis, modification/transformation and synthesis.

Javier Hernando received the MS and PhD degrees in Telecommunication Engineering from the Universitat Politècnica de Catalunya (UPC) in 1988 and 1993. He has been with the Department of Signal Theory and Communications, UPC, where he is now an Associate Professor. His research interests include robust speech analysis, speech recognition, speaker verification and localization, oral dialogue, and multimodal interfaces. He is the author of about two hundred scientific publications.

Notes

1 Generally, systems that use both source and filter parameters perform better than systems that use solely source parameters, when systems are evaluated by means of generic background models and without impostors who employ intentional voice mimicking techniques. Where a speaker recognition system utilises both source and filter parameters, the question arises whether either the source or the filter parameters are more vulnerable to intentional mimicking. In Lau et al. (2004), it turned out that the mimicking subjects, both with and without training in phonetics, found it easier to mimic the source parameters of the target speaker than the filter parameters. Another study showed, however, that a professional voice imitator from the entertainment industry was clearly able to approximate the filter parameters of a well-known target speaker (Zetterholm, 2006).

References

Atal, B.S. (1972) Automatic speaker recognition based on pitch contours. *Journal of the Acoustical Society of America* 52(6B): 1687–1698. December.

Boersma, P. and Weenink, D. (1992) *Praat Version 4.5.16.* Computer program, retrieved from http://www.praat.org.

Bonafonte, A., Höge, H., Kiss, I., Moreno, A., Ziegenhain, U., van den Heuvel, H., Hain, H.-U., Wang, X. S. and Garcia, M. N (2006) TC-STAR: specifications of language resources and evaluation for speech synthesis. In *Proceedings of the 5th International Conference on Language Resources and Evaluation* 311–314. Genoa, Italy, 22–28 May.

Carey, M. J., Parris, E. S., Lloyd-Thomas, H. and Bennett, S. (1996) Robust prosodic features for speaker identification. In *Proceedings of the ICSLP* vol. 3: 1800–1803. Philadelphia, PA, 3–6 October.

Duxans, H. (2006) *Voice Conversion Applied to Text-to-Speech Systems*. PhD dissertation. Universitat Politècnica de Catalunya, Department of Signal Theory and Communications, Barcelona.

Erro, D. and Moreno, A. (2005) A pitch-asynchronous simple method for speech synthesis by diphone concatenation using the deterministic plus stochastic model. In *Proceedings of the 10th International Conference on Speech and Computer SPECOM 2005* 321–324. Patras, Greece, 17–19 October.

Erro, D., Moreno, A. and Bonafonte, A. (2007) Flexible harmonic/stochastic speech synthesis. In *Proceedings of the 6th ISCA Workshop on Speech Synthesis*, *SSW6* vol. 6: 194–199. Bonn, Germany.

Farrús, M., Garde, A., Ejarque, P., Luque, J. and Hernando, J. (2006) On the fusion of prosody, voice spectrum and face features for multimodal person verification. In *Proceedings of the ICSLP* 2106–2109. Pittsburgh, 17–21 September.

Farrús, M., Hernando, J. and Ejarque, P. (2007) Jitter and shimmer measurements for speaker recognition. In *Proceedings of the Interspeech* 778–781. Antwerp, Belgium, 27–31 August.

Indovina, M., Uludag, U., Snelik, R., Mink, A. and Jain, A. (2003) Multimodal biometric authentication methods: a COTS approach. In *Proceedings of the Workshop on Multimodal User Authentication* 99–106. Santa Barbara, CA.

Kain, A. and Macon, M. (1998) Spectral voice conversion for text-to-speech synthesis. In *Proceedings of the ICASSP* 285–288. Seattle, WA, USA.

Künzel, H. J. (2000) Effects of voice disguise on speaking fundamental frequency. *Forensic Linguistics. The International Journal of Speech, Language and the Law* 7(2): 149–179.

Künzel, H. J., González-Rodríguez, J. and Ortega-García, J. (2004) Effect of voice disguise on the performance of a forensic automatic speaker recognition system. In *Proceedings of the Speaker and Language Recognition Workshop* 153–156. Toledo, Spain, 31 May – 3 June.

Lass, N. J., Trapp, D. S., Baldwin, M. K., Scherbick, K. A. and Wright, D. L. (1982) Effect of vocal disguise on judgments of speakers' sex and race. *Perceptual and Motor Skills* 54(3 Pt 2): 1235–1240.

Lau, Y. W., Tran, D. and Wagner, M. (2005) Testing voice mimicry with the YOHO speaker verification corpus. In *Knowledge-Based Intelligent Information and Engineering Systems* vol. 3684: 15–21. Berlin, Heidelberg: Springer.

Lau, Y. W., Wagner, M. and Tran, D. (2004) Vulnerability of speaker verification to voice mimicking. In *Proceedings of the International Symposium on Intelligent Multimedia, Video and Speech Processing*. Hong Kong, 20–22 October.

Laver, J. (1994) *Principles of Phonetics*. Cambridge: Cambridge University Press.

Lindberg, J. and Blomberg, M. (1999) Vulnerability in speaker verification. A study of technical impostor techniques. In *Proceedings of the Interspeech*. Budapest, Hungary, 5–9 September.

Markham, D. (1997) *Phonetic Imitation, Accent, and the Learner*. PhD dissertation. Lund University, Department of Linguistics and Phonetics, Lund.

Masuko, T., Tokuda, K. and Tobayashi, T. (2000) Imposture using synthetic speech against speaker verification based on spectrum and pitch. In *Proceedings of the ICSLP*. Beijing, China, 16–20 October.

Matrouf, D., Bonastre, J.-F. and Fredouille, C. (2006) Effect of speech transformation on impostor acceptance. In *Proceedings of the ICASSP*. Toulouse, France.

Perrot, P., Morel, M., Razik, J. and Chollet, G. (2009) Vocal forgery in forensic sciences. In M. Sorell (ed.) *Forensics in Telecommunications, Information and Multimedia* vol. 8 of *LNICST* 179–185. Berlin, Heidelberg: Springer.

Peskin, B., Navrátil, J., Abramson, J., Jones, D., Klusacek, D., Reynolds, D. A. and Xiang, B. (2003) Using prosodic and conversational features for high-performance speaker recognition. Report from JHU WS'02. In *Proceedings of the ICASSP* 792–795. Hong Kong, China.

Pittam, J. (1994) *Voice in Social Interaction: an interdisciplinary approach*. Thousand Oaks: SAGE Publications.

Reynolds, D. A., Andrews, W., Campbell, J., Navrátil, J., Peskin, B., Adami, A., Jin, Q., Klusacek, D., Abramson, J., Mihaescu, R., Godfrey, J., Jones, D. and Xiang, B. (2003) The SuperSID project: exploiting high-level information for high-accuracy speaker recognition. In *Proceedings of the ICASSP*, Hong Kong, China.

Rodman, R. D. (1998) Speaker recognition of disguised voices. In M. Demirekler, A. Saranli, H. Altincay and A. Paoloni (eds) *Proceedings of the Consortium on Speech Technology Conference on Speaker Recognition by Man and Machine: directions for forensic application* 9–22. Ankara, Turkey: COST250 Publishing Arm.

Shuy, R. W. (1995) Dialect as evidence in law cases. *Journal of English Linguistics* 23 (1/2): 195–208.

Stylianou, Y., Cappé, O. and Moulines, E. (1998) Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing* 6(2): 131–142.

Sullivan, K. P. H. and Pelecanos, J. (2001) Revisiting Carl Bildt's impostor: would a speaker verification system foil him? In *Proceedings of the 3rd International Conference on Audio- and Video-Based Biometric Person Authentication* vol. 2091: 144–149. Halmstad, Sweden.

Zetterholm, E. (2003) *Voice Imitation. A Phonetic Study of Perceptual Illusions and Acoustic Success*. PhD dissertation. Travaux de l'institut de linguistic de Lund 44. Lund University, Department of Linguistics and Phonetics, Lund.

Zetterholm, E. (2006) Same speaker – different voices. A study of one impersonator and some of his different imitations. In *Proceedings of the 11th Australian International Conference on Speech Science & Technology* 70–75. Auckland, New Zealand, 6–8 December.

Zetterholm, E., Blomberg, M. and Elenius, D. (2004) A comparison between human perception and a speaker verification system score of a voice imitation. In *Proceedings of the 10th Australian International Conference on Speech Science & Technology* 393–397. Sydney, Australia, 8–10 December.