Robustness of prosodic features to voice imitation

Mireia Farrús[1,2]*, Michael Wagner*[1]*, Jan Anguita*[1,2]*, Javier Hernando*[2]

[1]National Centre for Biometric Studies, School of Information Sciences and Engineering,
University of Canberra, Australia

[2]TALP Research Centre, Department of Signal Theory and Communications,
Technical University of Catalonia (UPC), Barcelona, Catalonia

{mfarrus, jan, javier}@gps.tsc.upc.edu, michael.wagner@canberra.edu.au

Abstract

Prosody plays an important role in the human recognition process; therefore, prosodic elements are normally used by impersonators aiming to resemble someone else. Since such voice imitation is one of the potential threats to security systems relying on automatic speaker recognition, and prosodic features have been considered for state-of-the-art recognition systems in recent years, the question arises as to what extent a mimicker is able to get close the prosodic characteristics of a target speaker. To this end, two experiments are conducted for twelve individual features in order to determine how a prosodic speaker identification system would perform against professionally imitated voices. The results show that the identification error rate increases for all the features except F0 range when the impersonators' modified voices are used instead of the impersonators natural voices. Moreover, it seems easier to copy prosody on the basis of a whole sentence than for a specific word.

Index Terms: speaker identification, voice imitation, prosody.

1. Introduction

Voice imitation is the reproduction of another speaker's voice and speech behaviour in order to pretend to be someone else [1]. A successful imitator has to be able to identify, select and imitate the most characteristic speech features of the target speaker. However, there are some organic differences between speakers, which cannot be changed, so that, when these differences are large, it may be difficult to achieve good imitations of another person's voice [2].

Speech features extracted by speech signal processing relate to the manner of sound generation in the larynx (*source*) on the one hand and to the acoustic filtering of the speech sounds in the vocal and nasal tracts (*filter*) on the other. Early automatic speaker recognition systems tended to use only the filter parameters, which relate to the physiology of the vocal tract and to the learnt articulatory configurations that shape the specific speech sounds [3]. More recently, some speaker recognition systems have begun also to use the source parameters, which relate mainly to the fundamental frequency and power (or perceived pitch and loudness) of the speech sounds and, in turn, to the prosody of the spoken phrases [4-6]. Generally, systems that use both source and filter parameters perform better than systems that just use source parameters, when systems are evaluated by means of generic background models and without impostors who employ intentional voice mimicking techniques.

Some recent studies have tested the vulnerability of automatic speaker recognition systems to intentional voice mimicking [7, 8]. Such vulnerability is of particular concern where these systems are used to control client access in applications such as telephone banking or other financial services. When both source and filter parameters are used, the question arises whether either the source or the filter parameters are more vulnerable to intentional mimicking. In [7], it transpired 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 [9].

In order to investigate further the question of how vulnerable automatic speaker recognition systems are to voice mimicking, the current study explores the ability of professional mimickers to approximate the source parameters and 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, twelve typical source-related parameters 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*). The experiments reveal, for each of the twelve source parameters, how much the professional imitator is able to shift the parameter away from his own voice and towards the target speaker's voice. In turn, these comparisons establish the robustness of the twelve source parameters against intentional voice mimicking by professionally trained impersonators.

2. Voice source and prosodic features

In addition to the acoustics of speech, humans tend to use several linguistic levels of information like the lexicon, prosody and phonetics to recognise others by their voice. These levels of information are normally related to learned habits or style, and they are mainly manifested in the dialect, sociolect or idiolect of the speaker.

Since these linguistic levels play an important role in the human recognition process, a lot of effort has been put into adding this kind of information to automatic speaker recognition systems. Recent works [4-6] have demonstrated that prosody helps to improve recognition systems based solely on filter parameters, supplying complementary information not captured in the traditional systems. Moreover, some of these parameters have the advantage of being more robust than spectral features to some common problems like noise, transmission channel distortion, speech level and distance between the speaker and the microphone.

However, there are other characteristics that may provide complementary information and could be of a great value for speaker recognition. Jitter and shimmer, for example, are measures of the cycle-to-cycle variations of fundamental frequency and amplitude, respectively, which have been largely used for the description of pathological voice quality. In [10] it was demonstrated that these features can improve a speaker verification system when treated as complementary features to spectral and prosodic parameters. The twelve features used in these experiments include:

Features related to word and segment duration:

log (number of frames per word)
length of word-internal voiced segments
length of word-internal unvoiced segments

Features related to fundamental frequency:

log (mean F_0)
log (max F_0)
log (min F_0)
log (range F_0)
F_0 *pseudoslope*: (last F_0 - first F_0)/(#frames)
F_0 absolute slope

Features related to cycle-to-cycle variations [10]:

jitter: cycle-to-cycle variation of F_0
shimmer (absolute): variability of the peak-to-peak amplitude in decibels
shimmer (apq3): three-point Amplitude Perturbation Quotient

Jitter and shimmer are not normally considered prosodic features; however, in this paper, all the features used will be referred to as prosodic features for simplicity.

3. Recognition experiments

3.1. Material

Two male professional imitators, who will be referred to by their initials, cc and qn, took part in our experiments. They have worked as professional imitators on radio and TV for more than 5 years. They both are Catalan native speakers and have a Central Catalan dialect.

Five well-known male politicians, who will be referred to by 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.

The recordings of the target speakers were taken from public radio interviews, made in local radio station 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 of the Technical University of Catalonia.

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 in order to copy the target voices as naturally as possible. The impersonator qn imitated the politicians JR, PM and XT, and cc imitated JB and JS. Table 1 shows imitators and target speakers together with the mean fundamental frequency of each speaker. Both impersonators recorded all the extracted sentences of each target speaker 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	F_0 (Hz)	Target	F_0 (Hz)
cc	121	JB	110
		JS	85
qn	110	JR	81
		PM	95
		XT	87

3.2. Experimental design

Both impersonators' voices (*i-natural* and *i-modified* voices) were recorded at the same time 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, it was not reliable to perform the recognition task with a conventional cepstral-based GMM method. In fact, the GMM system was tested and no identification errors were obtained. Therefore, only source- and prosody-related parameters were taken into account, since they seem to be more robust to mismatched recordings.

For each *i-natural*, *i-modified* and target voice, a vector of twelve source- and prosody-related features (listed in 3) was extracted to perform the identification experiments. The parameters were extracted using the Praat software for acoustic analysis [11], 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.

The identification experiments were divided into two sets: a text-independent and a text-dependent set. In the text-independent experiments, 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 closed set of two speaker models: the mimicker using his natural voice and the corresponding target speaker, on the basis of the single source parameter. Again for each individual parameter, a second experiment was conducted to establish the error rate of an identification system which tried to identify the target and *i-modified* voices from the same closed set of two speaker models: the impersonator speaking with his natural voice and his corresponding target speaker.

On the other hand, in the text-dependent experiments, the aim was to analyse how the i-*modified* voice differed from both i-*natural* and target voices. For each of the twelve features, the mean over all words of the i-*modified* voice was compared to the mean of the i-*natural* and target voices. Again for each feature, the i-modified voice was classified as the voice in which the analysed feature was closer.

Furthermore, in order to know to what extent the i-*modified* feature was shifted, an *Imitation Rate* was defined.

3.3. Text-independent identification

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 (always the same set of sentences) were used for training the models. The remaining sentences, together with the corresponding *i-modified* sentences, were used for testing. So, 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).

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. Finally, the fusion of all the individual features was performed in each experiment at the score level. The scores were normalised with the well-known z-score normalisation, which transforms the scores into a distribution with zero mean and unit variance, and they were then fused with the matcher weighting method, where each individual score is weighted by a factor proportional to the recognition rate [12].

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

The error rates are given for the whole prosodic systems, i.e. 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 are compared when using *k*=1 and *k*=3 in the k-Nearest Neighbour classification.

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

Training	1st NN		3rd NN	
	baseline	modified	baseline	modified
Five sentences	10.3	19.3	8.7	18.3
Ten sentences	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 biggest difference can be seen with the 1st Nearest Neighbour as classifier and 10 sentences used for training.

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

As it can be seen in the figure, the error rates increase in all the individual parameters except one: the range of the fundamental frequency (i.e. the difference between the maximum and minimum values of F_0), which remains steady, or even decreases in this case, in the modified system.

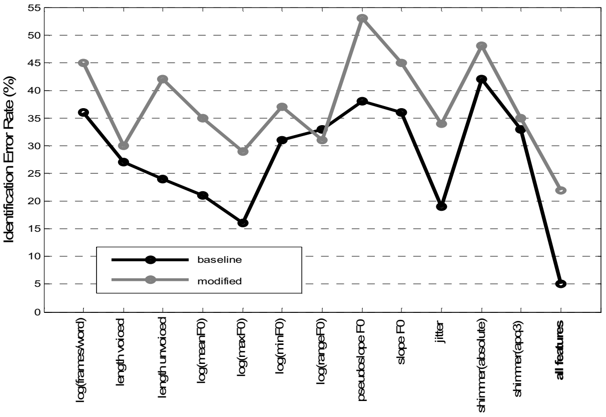

Figure 1: IER (%) for each prosodic feature (and fusion) using 1st NN and 10 sentences for training.

3.4. Text-dependent identification

The second part of this study analyses the similarity of the above-described prosodic and acoustic features between impersonator's *i-natural* and *i-modified* voices, and *i-modified* and target voices.

A first experiment established the IER of a system trying to identify the impersonator's *i-modified* from the i-*natural* and target voices. For every set of 20 different sentences corresponding to the i-*modified* voice, the mean over all words of each of the twelve parameters was computed. Then, each value was compared to the corresponding feature mean of the other two voices: target and *i-natural*. The comparison was always made between the same sentences, performing a total number of 100 tests (5 targets x 20 sentences). The identity of the i-*modified* voice was assigned to the speaker whose feature distance was closer (1st NN classifier).

A second experiment was performed in the same procedure as above, but in the basis of a specific word. The identification task was performed over each word of the sentence; then, the majority voting classifier was applied to the whole sentence at the decision level, assigning the identity of i-*modified* to the most voted model.

Furthermore, an *Imitation Rate* (IR) was defined as:

$$IR = \frac{\sum d_{i\text{-}natural}}{\sum d_{target}} \qquad (1)$$

where $\sum d_{i\text{-}natural}$ and $\sum d_{target}$ are the cumulative distances between the impersonator's *i-modified* and i-natural voices, and the *i-modified* and target voices, respectively. Note that IR > 1 signifies a "good" imitation. As in the previous experiments, the distances in the IR were computed in the basis of a whole sentence and for a specific word.

The obtained results are shown in Table 3. Both IER and IR values are also plotted in Figure 1 and 2, respectively, for each of the twelve individual parameters. The IER increases considerably when the identification is performed on the basis of whole sentences, so that the impersonator seems to be more successful when imitating the generic prosodic contour than the prosodic characteristics in every single word.

Table 3. IER (%) and "IR (imitation rate)" for prosodic and acoustic features (sentences and words).

Feature	sentences		words	
	IER	IR	IER	IR
log (#frames/word)	53	1.14	40	0.97
length voiced	46	0.77	27	0.85
length unvoiced	54	1.00	18	0.76
log (mean F_0)	34	0.52	13	0.67
log (max F_0)	34	0.62	27	1.00
log (min F_0)	32	0.55	8	0.63
log (range F_0)	24	0.44	7	0.67
F_0 pseudoslope	36	0.34	29	1.01
F_0 slope	29	0.53	21	0.69
jitter	30	0.60	15	0.67
shimmer (absolute)	49	0.93	26	0.89
shimmer (relative)	52	0.90	31	0.84
Fusion	24	-	7	-

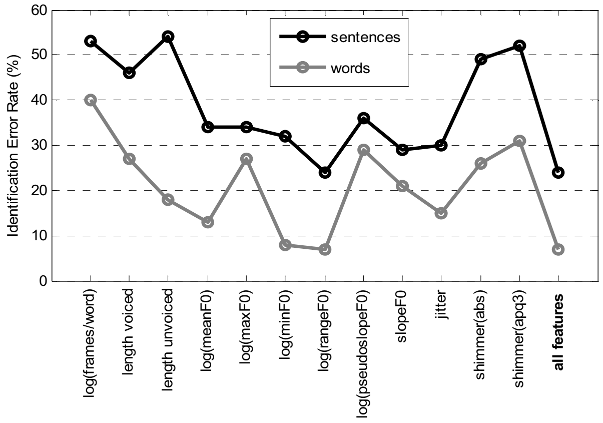

Figure 1: IER (%) using the analysis of prosodic features over sentences and over each word.

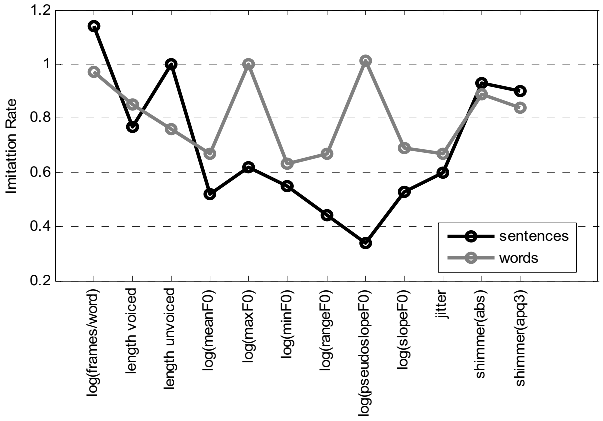

Figure 2: Imitation rate over sentences and words.

These results are also captured in the table and Figure 2 for four of the twelve features, where IR values are greater in the basis of whole sentences. However, IR values are smaller in the other features analysed. Moreover, in most cases the IR are below 1, which means that the impersonator's voice is still closer to his own voice. This suggests that, when the impersonator is identified as himself (i-*natural*), most of the features differ largely from the target; on the other hand, when the mimicker is identified as the target, his prosodic features are still relatively close to his own features.

4. Conclusions

A set of experiments was conducted, in which twelve prosodic and source-related features were used for speaker identification, and where a professional impersonator attempted to mimic a target voice. For each individual feature, the identification error rate *without* and *with* attempted impersonation was determined. For eleven of the twelve features, the IER with attempted impersonation increased, but for the F_0 range it remained almost unchanged. Fusing the twelve features resulted in an increase from an identification error rate of 5% to 22%. Another experiment showed that impersonators tended to imitate the prosody of a whole sentence rather than the prosodic characteristics for a specific word. These results show that the inclusion of prosodic and source-related 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. However, as the current database is small, the results should be interpreted with caution.

5. Acknowledgements

This work has been supported by the Spanish Government under grant AP2003-3598. The authors would like to thank Queco Novell, Cesc Casanovas, Jordi Ventura, Víctor Polo and Anna Pujol, who made possible the database recording.

6. References

[1] D. Markham, "Phonetic Imitation, Accent, and the Learner," *(PhD dissertation)*. Lund University, 1997.

[2] J. Laver, *Principles of phonetics*. Cambridge: Cambridge University Press, 1994.

[3] L.R. Rabiner, B.H.Juang, *Fundamentals of Speech Recognition*. Prentice Hall, Inc., 1993.

[4] B. Peskin et al., "Using prosodic and conversa-tional features for high-performance speaker recognition: Report from JHU WS'02," ICASSP, 2003.

[5] D.A. Reynolds et al., "The Super-SID project: exploiting high-level information for high-accuracy speaker recognition," ICASSP, 2003.

[6] M. Farrús et al., "On the Fusion of Prosody, Voice Spectrum and Face Features for Multimodal Person Verification," ICSLP, 2006.

[7] Y.W. Lau, M.Wagner, D.Tran, "Vulnerability of Speaker Verification to Voice Mimicking," ISIMP, 2004.

[8] Y.W.Lau, D.Tran, M.Wagner, "Testing Voice Mimicry with the YOHO Speaker Verification Corpus," *KES*, vol. 3684, *LNCS*. Springer, 2005, pp. 15-21.

[9] E.Zetterholm, "Same speaker - different voices. A study of one impersonator and some of his different imitations," 11th Australian International Conference on SST, 2006.

[10] M.Farrús et al.,"Jitter and Shimmer Measurements for Speaker Recognition," Eurospeech, 2007.

[11] Praat software website (Version 4.5.16): http://www.fon.hum.uva.nl/praat/.

[12] M. Indovina et al., "Multimodal Biometric Authentication Methods: A COTS Approach," Workshop on Multimodal User Authentication, 2003.