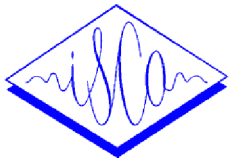
ISCA Archive
http://www.isca-speech.org/archive



Odyssey 2008: The Speaker and Language Recognition Workshop
Stellenbosch, South Africa
January 21-24, 2008

How vulnerable are prosodic features to professional imitators?

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]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

Voice imitation is one of the potential threats to security systems that use automatic speaker recognition. Since prosodic features have been considered for state-of-the-art recognition systems in recent years, the question arises as to how vulnerable these features are to voice mimicking. In this study, two experiments are conducted for twelve individual features in order to determine how a prosodic speaker identification system would perform against professionally imitated voices. By analysing prosodic parameters, the results show that the identification error rate increases for most of the features, except for the range of the fundamental frequency, which seems to be relatively robust against voice mimicking. When all twelve features are fused, the identification error rate increases from 5% between the target voices and the imitators' natural voices to 22% between the target voices and the imitators' impersonations.

1. Introduction

Speech signal processing extracts features from the speech signal, which 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 utilise solely the filter parameters, which relate - in a complex way - to the physiology of the vocal tract and to the learnt articulatory configurations that shape the specific speech sounds [1]. More recently, some speaker recognition systems have begun also to utilise 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 [2-4]. 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.

There are some recent studies which have tested the vulnerability of automatic speaker recognition systems to intentional voice mimicking [5, 6]. Such vulnerability is of particular concern where automatic speaker recognition is used to control client access in applications such as telephone banking or other financial services. 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 one of our earlier studies [5], 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 [7].

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, 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*).

Twelve source- and prosody-related parameters include the length of words and word segments, means, extrema and ranges of the fundamental frequency (F_0), and jitter and shimmer. For each of those parameters, a baseline speaker identification experiment was 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 source parameter. Then, a second experiment was conducted - again for each individual source 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 the single source parameter. It is the comparisons between the two experiments that 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 vulnerability of the twelve source parameters against intentional voice mimicking by professionally trained impersonators.

2. Voice imitation

Impersonation is defined as the reproduction of another speaker's voice and speech behaviour [8], and impersonators are persons who have the ability to pretend successfully to be someone else. A successful impersonator has to be able to identify, select and imitate characteristic 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 [9].

Several studies have been done to test the vulnerability of speaker recognition systems against imitations by human or synthetic voices. An experiment reported in [10] 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 [11, 12], where the impostor acceptance rate is increased by modifying the voice of an impostor in order to target a specific speaker.

The vulnerability of state-of-the-art speaker recognition to human imitations has been tested in some recent studies [5, 6], where the experiments showed that an impostor who knows a client speaker of the database with a similar voice to his own voice, could attack the system. On the other hand, some experiments reported in [13] showed that an automatic speaker verification system was more robust against an impostor who could closely imitate a target voice than those systems relying on human identification and verification.

3. 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 placed in adding this kind of information to automatic speaker recognition systems. Recent works [2-4] 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, which may provide complementary information and could be of a great value for speaker recognition. In [14] it was demonstrated that jitter and shimmer measurements can improve a speaker verification system as features complementary to spectral and prosodic parameters.

The features used in the current experiments are listed below. Although jitter and shimmer are not normally considered prosodic parameters, all the features below will be referred to, for simplicity, as prosodic features in this paper.

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 *pseudo-slope*: (last F_0 - first F_0)/(#frames)
- F_0 absolute slope

Features related to cycle-to-cycle variations [14]:

- 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

4. Recognition experiments

4.1. Material

Two male professional imitators, who will be referred to with their initials (cc and qn) took part in our experiments. They have been working 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 male well-known politicians, who will be referred to 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.

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.

Table 1: Mean F_0 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

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.

4.2. Experimental Setup

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. 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 [15], 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 closed 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 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. 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).

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 unitary variance:

$$x_{ZS} = \frac{a - mean(A)}{std(A)} \qquad (1)$$

(where a is the individual score and A is the set of scores to normalise), and they were then fused with the matcher weighting method, where each individual score is weighted by a factor proportional to the recognition rate [16].

4.3. Identification Results

The identification error rates (IER) obtained for both baseline and modified systems are presented in per cent 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.

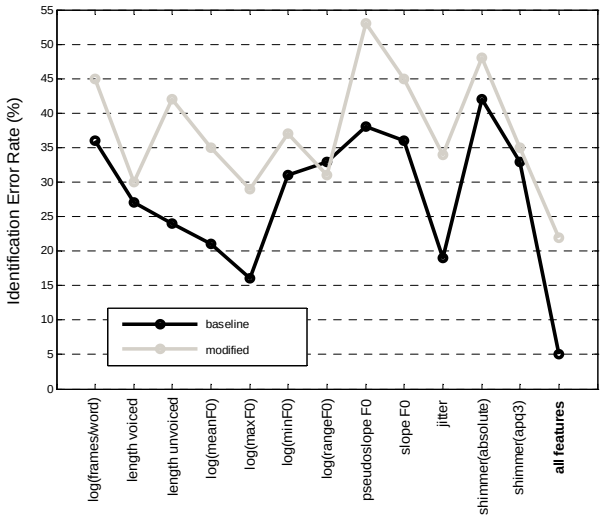


Figure 1: IER (%) for each prosodic feature (and fusion) using 1st NN and 10 sentences for training.

The identification error rates for each isolated feature are plotted in Figure 1, where the dark line corresponds to the 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 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.

5. 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 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 identification 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 F_0 range the identification error rate remained almost unchanged. Fusing the twelve features at the score level resulted in an increase from an identification error rate of 5% for target speakers against impersonators' natural voice to an identification error rate of 22% for target speakers against impersonators' modified voice. 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 database is small, the results should be interpreted with caution.

6. Acknowledgements

The authors would like to thank Queco Novell, Cesc Casanovas, Jordi Ventura, Víctor Polo and Anna Pujol, who made possible the database recording, and Ramon Cerdà for his help with the linguistic aspects.

7. References

[1] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1993.

[2] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. A. Reynolds, and B. Xiang, "Using prosodic and conversational features for high-performance speaker recognition: Report from JHU WS'02," ICASSP, 2003.

[3] D. A. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition," ICASSP, 2003.

[4] M. Farrús, A. Garde, P. Ejarque, J. Luque, and J. Hernando, "On the Fusion of Prosody, Voice Spectrum and Face Features for Multimodal Person Verification," ICSLP, Pittsburgh, 2006.

[5] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of Speaker Verification to Voice Mimicking," International Symposium on Intelligent Multimedia, Video and Speech Processing, Hong Kong, 2004.

[6] Y. W. Lau, D. Tran, and M. Wagner, "Testing Voice Mimicry with the YOHO Speaker Verification Corpus," in *Knowledge-Based Intelligent Information and Engineering Systems*, vol. 3684, *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer, 2005, pp. 15-21.

[7] E. Zetterholm, "Same speaker - different voices. A study of one impersonator and some of his different imitations," 11th Australian International Conference on Speech Science & Technology, Auckland, New Zealand, 2006.

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

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

[10] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification. A study of technical impostor techniques", Eurospeech, 1999.

[11] T. Masuko, K. Tokuda, and T. Tobayashi, "Imposture using Synthetic Speech Against Speaker Verification Based on Spectrum and Pitch," ICSLP, 2000.

[12] D. Matrouf, J.-F. Bonastre, and C. Fredouille, "Effect of speech transformation on impostor acceptance," ICASSP, Tolouse, France, 2006.

[13] K. P. H. Sullivan and J. Pelecanos, "Revisiting Carl Bildt's impostor: Would a speaker verification system foil him?," 3rd International Conference on Audio- and Video-Based Biometric Person Authentication, Halmstad, Sweden, 2001.

[14] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and Shimmer Measurements for Speaker Recognition," Eurospeech, Antwerp, Belgium, 2007.

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

[16] M. Indovina, U. Uludag, R. Snelik, A. Mink, and A. Jain, "Multimodal Biometric Authentication Methods: A COTS Approach," MMUA, Workshop on Multimodal User Authentication, Santa Barbara, CA, 2003.