INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Voice disguise vs. Impersonation : Acoustic and perceptual measurements of vocal flexibility in non experts

Véronique Delvaux[1,2], Lise Caucheteux[1], Kathy Huet[1], Myriam Piccaluga[1], Bernard Harmegnies[1]

[1]Institut de Recherche en Sciences et Technologies du Langage, UMONS, Belgium
[2]Fonds National de la Recherche Scientifique, Belgium
Veronique.Delvaux@umons.ac.be

Abstract

The aim of this study is to assess the potential for deliberately changing one's voice as a means to conceal or falsify identity, comparing acoustic and perceptual measurements of carefully controlled speech productions.

Twenty-two non expert speakers read a phonetically-balanced text 5 times in various conditions including natural speech, free vocal disguise (2 disguises per speaker), impersonation of a common target for all speakers, impersonation of one specific target per speaker. Long-term average spectra (LTAS) were computed for each reading and multiple pairwise comparisons were performed using the SDDD dissimilarity index.

The acoustic analysis showed that all speakers were able to deliberately change their voice beyond self-typical natural variation, whether in attempting to simply disguise their identity or to impersonate a specific target. Although the magnitude of the acoustic changes was comparable in disguise vs. impersonation, overall it was limited in that it did not achieved between-speaker variation levels. Perceptual judgements performed on the same material revealed that naive listeners were better at discriminating between impersonators and targets than at simply detecting voice disguise.

Index Terms: voice disguise, impersonation, vocal flexibility, LTAS

1. Introduction

Voice disguise is defined as the deliberate action of changing one's voice as a means to conceal identity. Impersonation is a particular category of voice disguise in which the aim of the impersonator is to falsify their identity, i.e. to sound like another person.

Detection of voice disguise is a crucial issue in forensic sciences, and many recent studies concern the design of automatic speaker recognition systems which can cope with voice disguise (e.g. [1, 2, 3]). In general phonetics, the study of voice disguise allows to address important issues such as the potential of human voice for flexibility and the extent of deliberate control over one's own speech productions.

Non electronic voice disguises may be classified based on the strategies used to achieve disguise, i.e. long-term deformation of supra-laryngeal structures (pinched nostrils, clenched jaw, lip protrusion, lowered velum, feigned lisp, etc.), overall changes in laryngeal function (falsetto, whisper, raised or lowered larynx, creaky voice, etc.), targeted modification of prosodic features (intonation, speech tempo, etc.), imitation of speech mannerisms considered as typical of specific groups of speakers (foreigners, elerdly people, rurals, etc.).

These diverse strategies may all result in changes in various aspects of oral productions: voice (the laryngeal source signal); segmental and suprasegmental features of speech; and overall voice quality (in the sense of the long-term acoustic quality of the speech signal resulting from vocal tract configuration, independently of segmental content). Particularly documented are modifications in fundamental frequency [1, 4, 5, 6, 7], formant frequencies mean and variability [1, 8, 9, 10], and speech rate [6, 10, 11]. On the other hand, the temporal organization of speech may be more resistant to vocal disguise [12, 13].

With plenty of different disguises possible, and the diversity of acoustic parameters potentially affected, it is difficult to accurately assess how *efficient* voice disguise is. That is, we know that human voice is flexible, but we know little of the extent of this flexibility and its limits. In the same vein, it is well documented that voice identification by both human listeners and automatic systems is affected by vocal disguise [3, 4, 10, 11, 14], but the exact acoustic correlates of this reduction in performances remain to be uncovered.

The aim of the present study is to assess the extent and limits of the changes deliberately made by non expert speakers (i.e.. non professional imitators) on their own voice as a result of free voice disguise and impersonation. Perceptual judgements by human listeners are compared with acoustic scores indexing the dissimilarity between overall speech productions. Task-induced deliberate modifications are assessed as a function of typical within-speaker variability, between-speakers variability, and in the case of impersonation, the acoustic similarity between the productions of the target and the imitator.

2. Material and methods

2.1. Study design and participants

The study comprised two parts, a production study in which naive speakers produced speech with and without voice disguise (including impersonation), and a perception study in which naive listeners perceptually judged a subset of these productions.

Participants in the production study were 22 French native speakers from Belgium and Northern France, 11 males (M1-M11), 11 females (F1-F11), averaging 22 years (SD: 2 years), who were documented for knowledge in linguistics and phonetics, musical and theater practices, as well as voice- and speech-related medical history. Participants in the perception

http://dx.doi.org/10.21437/Interspeech.2017-1080

study were a convenience sample of 23 French native speakers, 9 males, 14 females, aged 18 to 65.

The production study included two recording sessions of approximately 40 min, which consisted in multiple readings of the same phonetically-balanced French text [15] across various conditions. All recordings (16-bit, stereo, 44100 Hz) were made in a sound-attenuated room using a H5 Zoom© digital recorder including two matched unidirectional condenser microphones set at a 90 degree angle. In the first recording session, after a familiarization phase (silent reading), participants were instructed to read the text aloud (at least 5 times per condition): (i) as naturally as possible while avoiding reading errors (N1); (ii) by changing their voice so as not to be recognized, while not imitating another person/accent in particular, and avoiding the disguise to be detected by a third party (Da); (iii) using another strategy following the same instructions as in (ii) (Db).

In the second recording session, the participants read the text at least 5 times: (i) in their own, undisguised voice (N2); (ii) while sounding as much as possible like the voice A they just heard (Ia); (iii) the same as in (ii) with voice B (Ib). Voices A and B to be imitated were taken from the productions recorded in the first session. Voice A was the same for all speakers of a given gender. Voice A was selected as the voice which maximized dissimilarities with all other voices from the same gender in N1 based on SDDD (see below). Voice B was specific to each speaker. All B voices were selected so as to minimize the across-pair differences in within-pair dissimilarities based on SDDD.

In total, data collected in the production study amounted to 5 repetitions * 6 conditions (N1, Da, Db, N2, Ia, Ib) * 22 speakers, i.e. 660 readings.

The perception study comprised two tasks. In the first task, sentences extracted from the production study were presented by pairs and listeners had to specify on a 5-point Likert scale their dis/agreement with the statement that both sentences had been produced by the same speaker. In the second task, listeners had to specify on a 5-point Likert scale their dis/agreement with the statement that the single sentence they just heard had been produced with a disguised voice. The stimuli consisted in the middle sentence of the text (which segmental content most correlated with that of the whole text: $r = .91$; $p < .01$) from the first repetition in each condition for each speaker, for a total of 132 sentences.

2.2. Data processing and measures

For the acoustic analysis, 5 repetitions per condition per speaker were selected based on minimization of production errors. One long-term average spectrum (LTAS; frequency range: 22050 Hz; bandwidth: 25 Hz) was computed for each reading using Praat [16].

Multiple pairwise comparisons between LTAS were performed using the SDDD dissimilarity index, which is defined as the standard deviation of the level differences between the spectra under comparison. SDDD allows to summarize in one score multiple differences between two LTAS, with high sensitivity to difference in spectral shape and no sensitivity to overall intensity level [17]. Pairwise comparisons included:

- within-speaker within-condition comparisons, i.e. 10 pairwise comparisons between all 5 repetitions made in one speaker's undisguised voice (N1-N1, N2-N2) or disguised voice (Da-Da, Db-Db, Ia-Ia, Ib-Ib);

- within-speaker between-conditions comparisons, i.e. 25 pairwise comparisons between all 5 repetitions made in one speaker's undisguised voice vs. disguised voice (N1-Da, N1-Db, N2-Ia, N2-Ib);

- between-speaker within-condition comparisons, i.e. 25 pairwise comparisons between all 5 repetitions made in undisguised voice by an imitator and his target ($N1_{target}$-$N1_{imitator}$ for Ia pairs; $N1_{target}$-$N1_{imitator}$ for Ib pairs);

- between-speaker between-condition comparisons, i.e. 25 pairwise comparisons between all 5 repetitions made by an imitator (in his imitating voice) and his target (in his undisguised voice) ($N1_{target}$-$Ia_{imitator}$, $N1_{target}$-$Ib_{imitator}$).

The resulting average (over 10 or 25 comparisons) SDDD values constituted the dependent variables of the statistical analyses which were performed using SPSS©.

Perceptual judgements were coded as correct or incorrect responses accordingly, and as non responses when the listener selected the intermediate value of 3 ("I don't know") on the Likert scale. Proportions of correct, incorrect and non responses were averaged over all listeners and analyzed across pair types (task 1: speaker discrimination) or across conditions and speakers (task 2: detection of vocal disguise).

3. Results

3.1. Acoustic measures

3.1.1. Overall results

Overall results are illustrated in Figure 1 which plots error bars representing 95% CI of average SDDD as a function of comparison (all speakers included). A two-way analysis of variance was carried out, with SDDD as dependent variable and Gender (male; female) and Comparison as independent variables. Comparison was a 16-level variable of which 14 levels have been described above (N1-N1; N2-N2; Da-Da; Db-Db; Ia-Ia; Ib-Ib; N1-Da; N1-Db; N2-Ia; N2-Ib; $N1_{target}$-$N1_{imitator}$ for Ia pairs; $N1_{target}$-$N1_{imitator}$ for Ib pairs; $N1_{target}$-$Ia_{imitator}$; $N1_{target}$-$Ib_{imitator}$). The two remaining levels labelled "N1Inter" and "N2Inter" in Figure 1 represents average between-speaker variability in N1 and in N2. It was computed by carrying out 10 sets of 25 between-speaker within-condition pairwise comparisons, per speaker per session ($N1_{F1}$-$N1_{F2}$, $N1_{F1}$-$N1_{F3}$, $N1_{F1}$-$N1_{F4}$,... $N1_{F1}$-$N1_{F11}$) and averaging the resulting SDDD, thus giving an estimate of how different the productions of a given speaker were from those of all the other speakers of the same gender when all read in their undisguised voice.

The ANOVA indicated that there was a significant effect of Comparison on SDDD ($F(15,312)=50.787$, $p<.001$). Neither Gender, nor the interaction between Gender and Comparison, yielded significant differences in SDDD. Post hoc comparisons using the Scheffé's test revealed 5 homogeneous subsets among the 16 comparisons (Table 1). Interestingly, the first subset comprised all the within-speaker within-condition comparisons, the second subset all the within-speaker between-condition comparisons, and the three remaining subsets all the between-speaker comparisons.

In light of these results, a second two-way ANOVA was carried out, with SDDD as dependent variable and Speaker (22 levels) and Comparison Type (4 levels: within-speaker

within-condition; within-speaker between-condition, between-speaker within-condition; between-speaker between-condition) as independent variables. The ANOVA indicated that there was a significant main effect of both Comparison Type ($F(3,257)=443.509$, $p<.001$) and Speaker ($F(21,257)=3.031$, $p<.001$) on SDDD. The interaction effect was not significant. These results are illustrated in Fig.2. Post hoc comparisons using the Scheffé's test revealed that all 4 levels of Comparison Type were significantly different from each other in terms of SDDD (within-speaker within-condition: M=2.31, SD=.03; within-speaker between-condition: M=3.38, SD=.06, between-speaker within-condition: M=4.6, SD=.06; between-speaker between-condition: M=4.92, SD=.13).

Table 1: *Results of the post hoc Scheffé's tests (2-way ANOVA; Dependent Variable:SDDD; Independent Variables: Comparison & Gender)*: *Homogeneous subsets and associated means for Comparison*

Comparison	Subset				
	1	2	3	4	5
N2-N2	2.18				
N1-N1	2.19				
Db-Db	2.26				
Ia-Ia	2.32				
Da-Da	2.44				
Ib-Ib	2.49				
N2-Ia		3.31			
N1-Da		3.33			
N1-Db		3.42			
N2-Ib		3.47			
$N1_{target}$-$N1_{imitator}$ (Ia)			4.33		
N1Inter			4.35		
$N1_{target}$-$Ia_{imitator}$			4.47	4.47	
N2Inter			4.51	4.51	
$N1_{target}$-$N1_{imitator}$ (Ib)				5.21	5.21
$N1_{target}$-$Ib_{imitator}$					5.37

3.1.2. Free disguise

As illustrated on the lefthand side of Figure 1, overall the 22 speakers were efficient in changing their voice, in that the level of dissimilarity between productions made in their undisguised vs. freely disguised voice (N1-Da and N1-Db) was consistently and significantly higher than the inherent variability of their undisguised voice as measured in 5 successive repetitions of the same text in the same recording session (N1-N1). They were also able to be as sucessful with their first strategy as with their second strategy for disguisement (N1-Da did not significantly differ from N1-Db, see Table 1). However, from an acoustic perspective, the participants did not completely succeed in disguising their voice since the N1-Da and N1-Db dissimilarities remained significantly lower than the dissimilarities measured in the N1Inter condition. In other words, the 22 speakers were able to substantially alter their own voice, but not to reach a level of dissimilarity which would be compatible with typical between-speaker variation.

3.1.3. Impersonation

When the task was to imitate another person's voice, the same pattern as in free disguise emerged in the sense that N2-Ia and N2-Ib levels of dissimilarity lied in between N2-N2 and N2Inter (Figure 1 and Table 1). Recall that the targets for Ia and Ib were differentially selected in terms of dissimilarity between target and imitator (see $N1_{target}$-$N1_{imitator}$ for Ia pairs vs. $N1_{target}$-$N1_{imitator}$ for Ib pairs in Figure 1). Although the voice to be imitated was less similar to the participant's own voice in Ib than in Ia, overall the speakers were equally able to differ from their undisguised voice in Ia and Ib. However, dissimilarities did not significantly decrease between target and imitator as a result of imitation: there was no significant difference in SDDD between $N1_{target}$-$N1_{imitator}$ (Ia) and $N1_{target}$-$Ia_{imitator}$, nor between $N1_{target}$-$N1_{imitator}$ (Ib) and $N1_{target}$-$Ib_{imitator}$ (Table 1). In other words, the participants were able to significantly alter their own voices when imitating another individual, but in doing so they did not move closer to the specified target, at least in terms of LTAS.

3.2. Perceptual measures

Overall, perceptual results showed that listeners were less sucessful at detecting disguised voices when audio samples were presented one at a time (task 1: 57,36% of correct responses overall) than at discriminating between speakers based on pairs of audio samples (task 2: 78,02% of correct responses overall).

In task 1, almost one fourth of the responses were "I don't know" (23,26%), while most of the errors (4/5 out of 19,38%) were cases in which a disguised voice (Da, Db, Ia, Ib) was incorrectly judged as non disguised.

In task 2, non response and error rates respectively amounted to 14,29% and 7,69%. There was no error when the two sentences of the pair had been produced by the same speaker in his undisguised voice (N1-N2 pairs). Errors first concerned pairs of sentences produced by the same speaker in his/her undisguised vs. disguised voice (N1-Da; N1-Db; N2-Ia; N2-Ib), which were incorrectly judged as belonging to two different speakers. In a lesser proportion, other errors concerned pairs of sentences produced by two different speakers, which were incorrectly judged as uttered by the same person. However, these errors were almost as frequent in $N1_{target}$-$N1_{imitator}$ pairs than in $N1_{target}$-$Ia_{imitator}$ or in $N1_{target}$-$Ib_{imitator}$ pairs. This means that the listeners had difficulties to discriminate between the speakers in some pairs, whether or not one was trying to sound like the other.

Overall, the perceptual judgements collected in Task 2 were consistent with the acoustic analysis performed above. The patterns of perceptual errors were compatible with speakers being able to modify their own voice, but usually not to the point of being confused with someone else, even less with a specific individual. However, no significant correlation could be found between SDDD values and perceptual scores in task 2, whether overall or for specific comparison types.

4. Conclusion

The aim of the present study was to study the changes deliberately made by non expert speakers on their own voice comparing free voice disguise and impersonation. The methodology used in the present paper allowed to assess the extent and limits of the observed changes. First, the acoustic analysis showed that all speakers were able to change their voice beyond self-typical natural variation, whether in attempting to simply conceal their identity or to impersonate a specific target. Results of the first perceptual task revealed that disguised voice were not easily detected as such when no

reference point was provided to the listeners. However, the magnitude of disguise-related acoustic changes was limited in that it did not reach between-speaker variation levels (as evidenced in the acoustic analysis), even if in some instances it was sufficient to deceive naive listeners (as evidenced in the second perceptual task). Second, our speakers were far better at free disguise than at impersonation. Although the participants were able to significantly alter their own voices when imitating another individual, in doing so they did not move closer to the specified target, either in acoustic or perceptual terms. Implications of these findings for forensic sciences are mutiple and will be discussed at the conference. Perspectives of the present study include refined attempts to relate acoustic measures (based on LTAS and other acoustic parameters such as MFCC and/or voice source parameters) with perceptual scores, as well as a detailed, qualitative analysis of the results as a function of individual profiles (e.g. with respect to theater and musical practices) and self-reported disguise strategies.

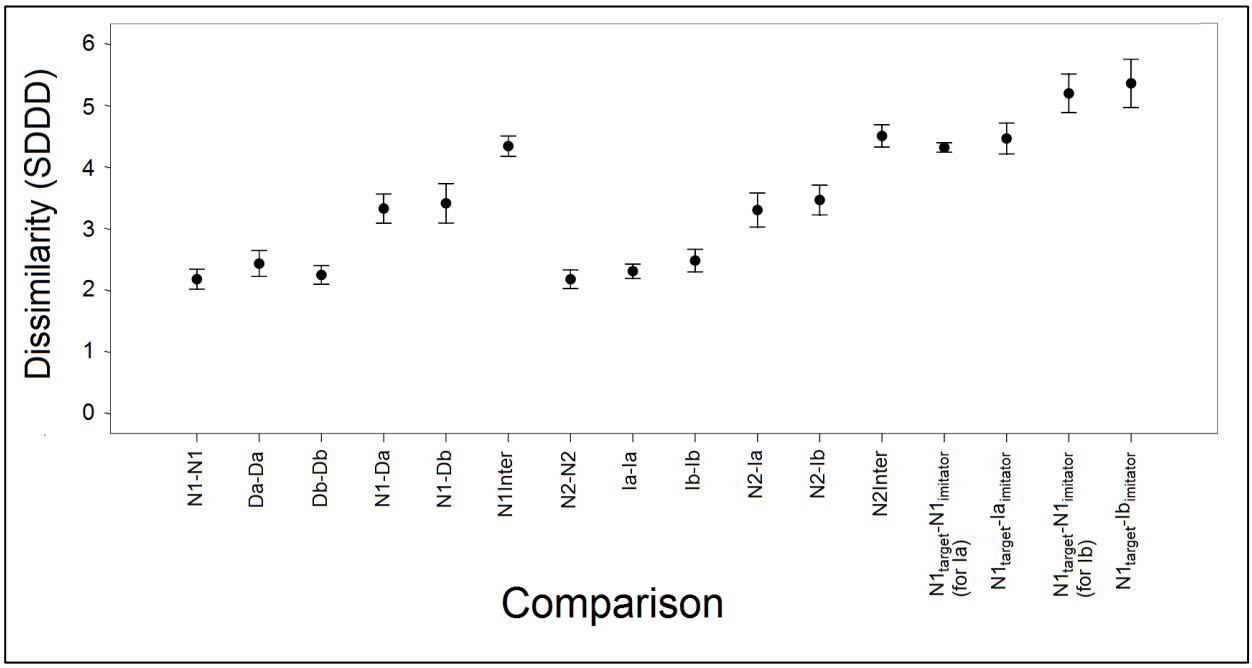

Figure 1: *Dissimilarity (SDDD) as a function of Comparison (16 levels) over the 22 speakers. Error bars represent 95% confidence intervals. See text for details.*

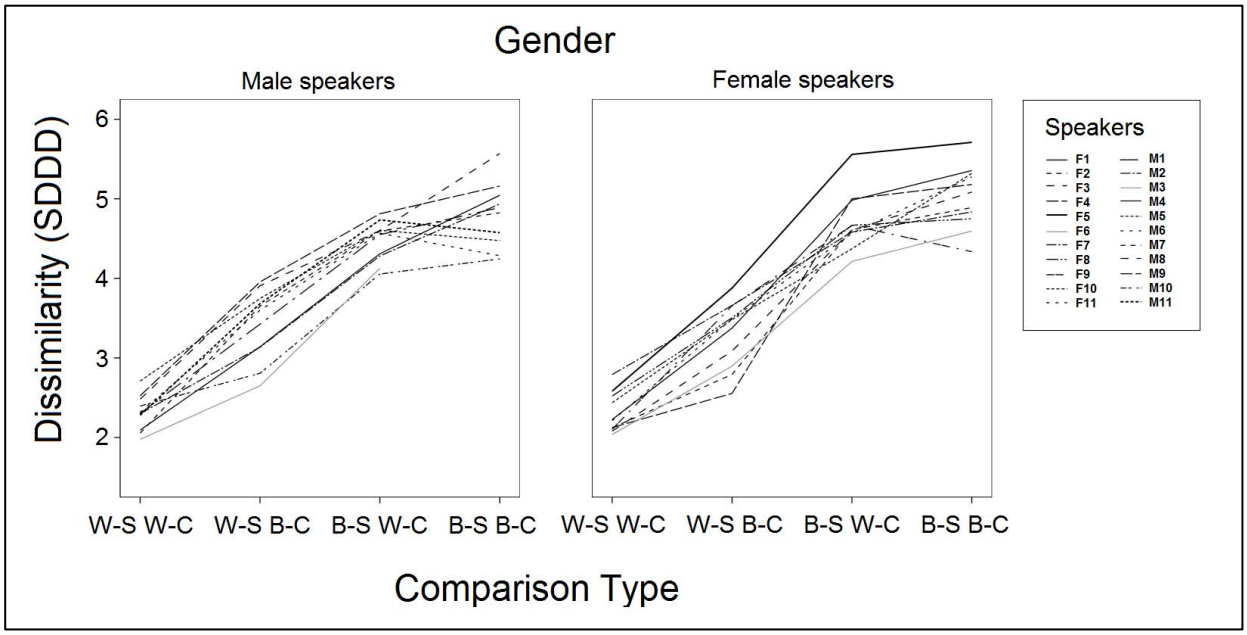

Figure 2: *Dissimilarity (SDDD) across Comparison Type: W-S W-C: Within-speaker within-condition, W-S B-C: Within-speaker between-condition, B-S W-C: Between-speaker within-condition, B-S B-C: Between-speaker between-condition. Left: individual male speakers, right: individual female speakers*

5. References

[1] Perrot, P., and Chollet, G. (2012). "Helping the Forensic Research Institute of the French Gendarmerie to identify a suspect in the presence of voice disguise or voice forgery," in A. Neustein, H.A. Patil (Eds.), *Forensic Speaker Recognition* (pp. 469-503). Berlin, Allemagne : Springer-Verlag Berlin Heidelberg. Doi : 10.1007/978-1-4614-0263-3

[2] Zhang, C., and Tan, T. (2007). "Voice disguise and automatic speaker recognition," *Forensic Science International, 175*, 118-122. Doi : 10.1016/j.forsciint.2007.05.019

[3] Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., and Li, H. (2014). "Spoofing and countermeasures for speaker verification : A survey," *Speech Communication, 66*, 130-153. Doi : 10.1016/j.specom.2014.10.005

[4] Wagner, I., and Köster, O. (1999). "Perceptual recognition of familiar voices using falsetto as a type of voice disguise," *Proceedings of the 14th International Congress of Phonetic Sciences*, 1381–1384.

[5] Zetterholm, E. (2006). "Same speaker, different voices : A study of one impersonator and some of his different imitations," *Proceedings of the 11th Australasian International Conference on Speech Science and Technology SST2006*. University of Auckland, Auckland, New Zealand. Paper 41.

[6] Amin, T.B., Marziliano, P., and German, J.S. (2012). "Nine voices, one artist : Linguistic and acoustic analysis," *Proceedings of IEEE International Conference on Multimedia and Expo* 2012 (ICME2012), pp. 450-454.

[7] Gomes, M.L., and Kremer, R.L. (2015). "Fundamental frequency and the strategies of disguise : A comparison of harsh voice and lip protrusion in male and female speakers," Retrieved from http://media.leidenuniv.nl/legacy/kremerrobinson.pdf

[8] Ashour, G., and Gath, I. (1999). "Characterization of speech during imitation," *Proceedings of Eurospeech 1999*.

[9] Kitamura, T. (2008). "Acoustic analysis of imitated voice produced by a professional impersonator," *Proceedings of Interspeech 2008*.

[10] Amin, T.B., Marziliano, P., and German, J.S. (2014). "Glottal and vocal tract characteristics of voice impersonators," *IEEE Transactions on Multimedia, 16*, 668-678. Doi : 10.1109/TMM.2014.2300071

[11] Revis, J., De Looze, C., and Giovanni, A. (2013). "Vocal flexibility and prosodic strategies in a professional impersonator," *Journal of Voice,* 27, 524.e23- 524.e31. Doi : 10.1016/j.jvoice.2013.01.008

[12] Dellwo, V., Huckvale, M., and Ashby, M. (2007). "How is Individuality expressed in voice ? An introduction to speech production and description of speaker classification," In C. Müller (Ed.), *Speaker Classification I* (pp. 1-20). Berlin, Allemagne : Springer-Verlag Berlin Heidelberg. Doi : 10.1007/978-3-540-74200-5_1

[13] Leemann, A., and Kolly, M.-J. (2015). "Speaker-invariant suprasegmental temporal features in normal and disguised speech," *Speech Communication, 75*, 97-122. Doi : 10.1016/j.specom.2015.10.002

[14] Perrot, P., Aversano, G., and Chollet, G. (2007). "Voice disguise and automatic detection : Review and perspectives," In Y. Stylianou, M. Faundez-Zanuy, A. Esposito (Eds.), *Progress in Nonlinear Speech Processing* (pp. 101-117). Berlin, Allemagne : Springer-Verlag Berlin Heidelberg. Doi : 10.1007/978-3-540-71505-4_7

[15] Harmegnies, B. (1988). *Contribution à la caractérisation de la qualité vocale : Analyses plurielles de spectres moyens à long terme de la parole* (Unpublished doctoral dissertation). Université de Mons, Mons, Belgique.

[16] Boersma, Paul (2001). "Praat, a system for doing phonetics by computer," *Glot International* **5:9/10**, 341-345.

[17] Harmegnies, B. (1988). "SDDD : A new dissimilarity index for the comparison of speech spectra," *Pattern recognition letters, 8*, 153-158.