A COMPARISON OF INDICES OF DIFFERENCE AND SIMILARITY, BASED ON LTASS AND TESTED ON VOICES IN REAL FORENSIC CASE AND IN CONTROLLED CONDITIONS

Gordana Varošanec-Škarić and Jordan Bićanić

Faculty of Humanities and Social Studies, Dpt. of Phonetics, Univeritiy of Zagreb
gvarosan@ffzg.hr, jbicanic@ffzg.hr

ABSTRACT

This study compares the difference of indices of difference and similarity between voices in real case and in the controlled group. The LTASS has been set for the range of 800 and 3500 Hz. This result is based on the comparisons of indices of difference (SDDD: Standard Deviation of Difference Distribution) and similarity (R) in speech recordings of standardized text read by the controlled group. The real case was 6 male voices: 5 Croatian and 1 Albanian. The controlled group was 30 male and 35 female speakers (all Croatian). To enable data comparison, average values of indices were calculated for the same and different speakers on the basis of different parts of the spectrum: from 0 to 10 kHz and filtered voices from 0.8 to 3.5 kHz. The results of t-test have shown that the groups differ significantly, and as expected the difference was greatest respectively in the group of male voices recorded in the studio (0-10 kHz: $p<0.001$, $t=46.17$); filtered studio voices ($p<0.001$); and real case ($p<0.01$). It is of methodological importance that in real case the difference of SDDD indices between different and same pairs of voices was significant ($p<0.01$), as well as the difference of similarity indices ($p=0.01$).

Keywords: forensic phonetics, SDDD index, similarity index, LTASS.

1. INTRODUCTION

The background for this study is the forensic phonetics' viewpoint that in speaker identification, auditory perceptual evaluation is more valid than acoustic measuring [7, 12]. Because acoustic measuring cannot achieve 100% certainty in speaker identification, due to different speech contexts, it will be examined as an additional method. The 2 chosen methods were: the long-term average spectrum of speech (LTASS) and acoustic-statistical procedures that have the best record in speaker identification based on voice and sounds - SDDD index (Standard Deviation of Difference Distribution) and similarity index, i.e. inter-correlation coefficient (R) between spectra [3]. The acoustic-statistical procedures were also used in other phonetic research comparing different language speakers [1, 4, 5]. If phonetic auditory speaker identification is the essential procedure, the results of statistical-acoustic procedures are expected to show the difference between groups of same and different voices. We can agree with the statement by Rodman et al. [11] who claim that it is good for identification to be text independent. The starting point of this research is testing voices in real forensic cases, analyzing different longer speech contexts, because these results will provide data for comparison.

2. PROCEDURE

2.1. Preliminary examination

5619 compressed and cleaned recordings of real cases of conversations over GSM mobile phones (independent of the first and second author) were listened to before beginning the procedure of speaker identification. This involved 1 CD with "familiar speakers" and 5 CDs with "unfamiliar speakers". Harrison [6] talks about the need for the filtration of forensic audio recordings because of GSM interference. The pairs of voices were assembled in random order. Then the expert auditory speaker identification was conducted, according to the usual procedure in forensic phonetics AP-SPID (Aural-Perceptual Speech Identification) [7, 10]. Three phoneticians (including the first author) evaluated the 43 pairs of voices. Corresponding identities were established for six male voices.

In preliminary examination, the LTASS (program AS - Average Spectrum) was made for voices in real case on the basis of the whole area of transmission, and recorded on CD – from about

300 Hz to above 4 kHz. Telephone transmission affects mostly F1 of most of Croatian vowels. This is not the case with Croatian vowel /a/, in which the average F1 is around 800 Hz. The difference of timbre was only noticeable above 800 Hz, while above 4 kHz the spectrum plummeted. Therefore, it has been determined that the area from 800 to 3500 Hz should be the variable for the calculation of R and SDDD.

2.2. Material

The real cases were recorded in 2004 (recordings of speech over GSM mobile phones transferred to CD). Six male voices (5 Croatian and 1 Albanian speaking Albanian and Croatian) in different speech contexts have been edited: each voice with itself and different voices between themselves (43 pairs). As there was enough recorded material, long parts of speech, around 40 to 60 s, were used for creating the LTASS of real cases.

Control groups are recorded in two recordings (30 male and 35 female voices – all Croatian) of the same standardized text around 60 s.

2.3. Recording in controlled conditions

For the controlled conditions part of the experiment, 30 male and 35 female voices were recorded twice, with an interval of one month in-between, according to the determined procedure: in a well equipped phonetic studio at the same distance (30 cm) from the microphone. The microphone (AKG C414 ULS) was connected to the mixer that is used to control the input intensity. It was further connected to the outer part of the sound card (Soundscape iBox SS8IO-3 audio interface), which is an 8-channel analog/digital converter that is connected to the PCI sound card Soundscape Mixtrime PCI 16 which is in turn connected to the computer over the PCI slot.

2.4. Acoustic variables and acoustic-statistical procedures in main examination

In this way, all frequencies below 800 and above 3500 Hz were removed, and low and high areas of spectrum in the LTASS were avoided. This was done in order to get better results and eliminate the possible negative influence of removed spectral areas, on that subject: [2, 8, 9]. The usual band-pass mentioned for the latest GSM mobile phones is that of transmission channel from 350 to 3400 Hz [8], that was in accordance with the sample of real cases. In order to be able to compare the results, voices recorded in experimental conditions were also filtered in the same area as real cases, but for the control group the ideal variable 0-10 kHz was determined as well.

The main acoustic-statistical procedures of SDDD index and R similarity index are calculated on the basis of comparison of LTASS curves in 1024 points in the area between 0 and 10 kHz and in 277 points, in the area between 800 and 3500 Hz, for real cases and filtered voices recorded in controlled conditions. As the difference for the latter area is 2700 Hz, a special program is adapted for that area. It calculates SDDD index and similarity index on the basis of the comparison of 277 points between spectra. In theory, SDDD would be 0 (zero) when LTASS curves would be almost parallel, and the largest standard deviation would appear in the case when there would be no points of agreement. R – Similarity index, i.e. coefficient of inter-correlation informs us on the amount of co-variation of the two spectra.

Afterwards, t-test was used to examine the difference between groups of pairs of same and different voices, each group of speakers separately: real case and voices recorded in the studio.

3. RESULTS AND DISCUSSION

3.1. SDDD and R

The results on the basis of real forensic cases show that the average value of SDDD for same speakers in different context is 1.83: and for different speakers is 2.76. The average value of R for the same speakers is 0.92 and for different speakers 0.84. For both indices, ranges were significantly larger for different pairs of voices than for the group of same pairs of voices (Table 1). In expert evaluation, same speakers are evaluated with 96.02% of agreement, with average ranges from 89.00 to 100%. Different speakers are evaluated with 30.72% of agreement, with average ranges from 9.00 to 69.67%. It is interesting that the best results for identifying the same person were achieved for the male speaker who was identified as the leader of the group. On the basis of different recordings of speech, R of 0.97 shows the high similarity in the group of 6 identified voices and lower SDDD of 1.30 (Figure 1). The greatest difference was found in the male voice with R = 0.81 and SDDD = 3.42 (Figure 2). This can be explained by the fact that the leader was always speaking with the similar voice quality, paralinguistically always authoritative, in different social environment.

Figure 1: LTASS of male voice with lower value of SDDD index and higher value of similarity index (R) in two recodings (real case)

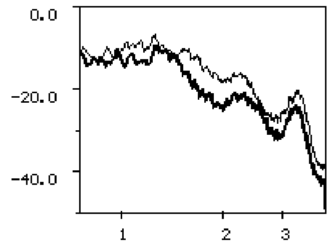
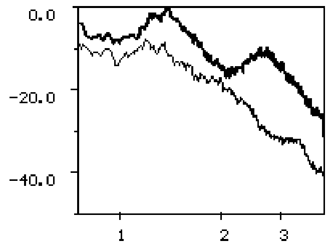


Figure 2: LTASS of male voice with higher value of SDDD index and lower value of similarity index (R) in two recodings (real case)



The fact that the leader was paralinguistically always authoritative, influenced the constant timbre which was showing the characteristic similar LTASS with prominent F4. This is because the dynamic characteristics of voice were similar as well: loudness was always between middle and high, so the pitch F0 was always about equal, as well as the intensity and shape of F4 and similar pronunciation and prosody. The voice with the greatest difference (younger male voice, lower hierarchy inside the group) changed depending on the different conditions: from very quiet speaking because of the fear of eavesdropping to very loud speaking in dangerous situations of ?police escaping? The range of SDDD in real cases for different speakers was from the lowest value of 1.76 to 5.38. The range of R was from 0.42 to 0.97. The range was smaller for the same pairs of voices: SDDD from 1.05 to 3.42 and R from 0.81 to 0.98. These data also prove that it is extremely important that the acoustic analysis is preceded by the phonetic auditory perceptional identification of speakers, and must be an essential procedure of speaker identification. This is because acoustic analysis involves all aspects of voice quality, constant timbre, and dialectal pronunciation, prosody, pronunciation that is for a certain speaker recognizable in different speech context and environment. In LTASS analysis, constant characteristics of voice, such as timbre itself, predominate. Therefore the LTASS and acoustic-statistical procedures can be useful additional methods of voice comparison, supplementary to the procedure of auditory speaker identification.

The largest significant differences between the same and different pairs of voices when speaking in a neutral tone in studio recording are obtained for male and female voices in direct recording on the basis of the LTASS for variable 0 – 10 kHz (Table 3). However the comparison with the real case data on the basis of filtrated speech is also interesting. The average value of SDDD for different male (2.68) and female (2.56) pairs of voices are slightly lower than the average value for the real cases. The average value of similarity index for different male voices is also comparable (0.83). The average index of similarity for same male pairs of voices (0.96) is slightly higher than for real cases, while the SDDD for the same pairs is higher in real case, as expected, but still comparable with filtrated voices (R = 1.07 for women and 1.10 for men). The range of SDDD is slightly higher in real cases, but is also comparable with filtrated male voices (from 1.13 to 5.35; Table 2). Although lower, the maximum value of SDDD for male voices is also above 5, and the maximum values of same male pairs of voices are also comparable (both values go above 3), while the range was smaller for the group of same female voices, who have smaller range of similarity index (from 0.96 to 1).

3.2. T-test results for the three groups of voices.

Concerning the different speech contexts, background noise and the conditions of GSM transmission in real cases, it is encouraging to have significant difference of SDDD indices between groups of same and different pairs of voices ($p<0.001$) and lower significant difference of inter-correlation coefficients between the groups ($p=0.01$; Table 1). The difference was significant for filtrated pairs (variable 800-3500 Hz) of male and female different and same voices according to the index of similarity and in relation to index of difference ($p<0.001$; Table 2). T-test has shown that the most significant differences between groups of different and same pairs of unfiltered voices in two recordings of the same text, arise in the variable 0-10 kHz, and that the similarity index of male voices has the greatest significant difference ($p<0.001$, $t = 46.17$; table 3).

Table 1: Average values and ranges of similarity indices (R) and SDDD indices, and t-test results in the variable between 0.8 and 3.5 kHz (real case).

		R	SDDD
Pairs of different speakers	x̄	0.84	2.76
	Min	0.42	1.76
	Max	0.97	5.83
Pairs of same speakers	x̄	0.92	1.83
	Min	0.81	1.05
	Max	0.98	3.42
T-test	p	0.01	<0.01
	t	2.75	3.82

Table 2: Average values and ranges of similarity indices (R) and SDDD indices, and t-test results in the variable between 0.8 and 3.5 kHz (controlled conditions).

		R		SDDD	
		Female	Male	Female	Male
Pairs of different speakers	x̄	0.91	0.83	2.56	2.68
	Min	0.75	0.58	0.76	1.13
	Max	0.98	0.97	5.21	5.35
Pairs of same speakers	x̄	0.98	0.96	1.07	1.10
	Min	0.96	0.87	0.38	0.12
	Max	1.00	1.00	1.89	3.29
T-test	p	<0.001	<0.001	<0.001	<0.001
	t	24.38	19.26	18.7	14.94

Table 3: Average values and ranges of similarity indices (R) and SDDD indices, and t-test results in the variable between 0 and 10 kHz (controlled conditions)

		R		SDDD	
		Female	Male	Female	Male
Pairs of different speakers	x̄	0.92	0.92	3.98	4.27
	Min	0.79	0.81	1.85	1.77
	Max	0.98	0.98	7.31	7.86
Pairs of same speakers	x̄	0.99	0.99	1.08	1.16
	Min	0.98	0.98	0.28	0.26
	Max	1.00	1.00	1.93	2.00
T-test	p	<0.001	<0.001	<0.001	<0.001
	t	38.01	46.17	31.5	36.48

4. CONCLUSION

It is interesting that for the real forensic cases group, the similarity indices are highest and SDDD lowest for people at the top of the hierarchy (i.e. leaders) and vice versa for the people at the bottom of hierarchy.

Acoustic-statistical procedures of SDDD and R obtained on the basis of the LTASS have shown reliable for voice distinguishing: same and different pairs of voices statistically differ significantly in real case despite the different speech context and paralinguistic factors. Relatively good results for real case can be assigned to the adequate amount of speech, because the LTASS gives the information on the constant timbre. Further on, the variable 800-3500 Hz has turned out to be very good in the comparison of pairs of voices, and it shows that the recorded and cleaned conversation over GSM mobile phones is adequate for perceptual and acoustic speaker identification.

5. REFERENCES

[1] Bruyninckx, M., Harmegnies, B., Llisteri, J., Poch-Olive, D. 1994. Language-induced voice quality variability in bilinguals. *Journal of Phonetics* 22, 19–31.

[2] Foulkes, P., Barron, A. 2000. Telephone speaker recognition amongst members of a close social network. *Forensic linguistics* 7, 2, 180–198.

[3] Harmegnies, B. 1995. Contribution à la Caractérisation Acoustique des Sigmatismes – Étude de deux Indices Acoustico-Statistiques. In Braun, A., Köster, J-P. (eds), *Studies in Forensic Phonetics,* Trier: WVT Wissenschaftlicher Verlag Trier, 56-66.

[4] Harmegnies, B., Landercy, A. 1985. Language Features in the Long-Term Average Spectrum. *Revue de Phonetique Apllique* 73–74–75, 69–79.

[5] Harmegnies, B., Landercy, A., Bruyninckx, M. 1987. An experiment in inter – language recognition using SDDD index. *Proc. 11th ICPhS* Tallin, Vol. 2, 241-244.

[6] Harrison, Ph. 2001. GSM interference cancellation for forensic audio: a report on work in progress. *Forensic Linguistics* 8, 2, 9–23.

[7] Hollien, H. 2002. *Forensic Voice Identification*. San Diego: Academic Press.

[8] Künzel, H. J. 2001. Beware of the 'telephone effect': the influence of telephone transmission on the measurement of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics* 8, 1, 80–99.

[9] Nolan, F. 2002. The 'telephone effect' on formants: a response. *Forensic Linguistics* 9, 1, 74–82.

[10] Nolan, J. F. 1983. *The Phonetic Basis of Speaker Recognition*. Cambridge: Cambridge University Press.

[11] Rodman, R., McAllister, D., Bitzer, D., Cepeda, L., Abbitt, P. 2002. Forensic speaker identification based on spectral moments. *Forensic Linguistics* 9, 1, 22–43.

[12] Varošanec-Škarić, G. 2004. Aural-perceptual approach to speaker indentification in forensic phonetics. *Knjiga sažetaka petog znanstvenog skupa Istraživanja govora* Zagreb, 108.