ANALYSIS OF SPECTRAL FEATURES' CHANGE FOR SPEAKER COMPARISON

Tekla Etelka Gráczi[1], Valéria Krepsz[1, 2], Anna Huszár[1], Bettina Száraz[1], Andrea Deme[3], Kornélia Juhász[1, 3], Alexandra Markó[1, 4]

[1]Hungarian Research Centre for Linguistics, [2]Humboldt-Universität zu Berlin, [3]Eötvös Loránd University, [4]SSNS Institute for Expert Services

graczi.tekla.etelka@nytud.hu, krepsz.valeria@nytud.hu, huszar.anna@nytud.hu, szaraz.bettina@nytud.hu, deme.andrea@btk.elte.hu, juhasz.kornelia@nytud.hu, marko.alexandra.phd@gmail.com

ABSTRACT

Medium-term changes of speech parameters are understudied, while forensic tasks often require the comparison of non-contemporary speech samples. The present study compares segmental features of the same speakers' speech recorded a decade apart.

F_1–F_4 values of five vowels and four spectral moments of five voiceless obstruents were measured in Hungarian reading aloud tasks. The ratios of the mean values were compared between-sessions, and the possible group-level change was tested by LMM. The relevance of the spectral features in speaker verification was analysed using LDA within the first and between the two recording sessions.

Most spectral features showed high intraspeaker variability in terms of their change between the sessions. The LDA-results showed that the sibilants may have similar relevance in within- and between sessions speaker verification, while vowels though having the highest relevance in within-session verification, lose on this in the case of between-session comparisons. Stops had the lowest relevance.

Keywords: spectral changes, speaker comparison, vowels, voiceless obstruents

1. INTRODUCTION

The most common task in forensic speaker identification involves acoustic and auditory comparison of several voice samples. Traditionally, analyses primarily focus on voice parameters (mainly f_0 and voice quality) and vowel formants [1, 2]. However, in studies of English spectral features of consonants have also been involved in forensic context recently [3]. Study of spectral parameters of /m, n, ŋ, l, s/ [4] found high speaker-specificity effect of /m/ and /s/. Earnshaw [5] studied voiceless plosives (/t/, /k/ and /t/ & /k/combined) in terms of VOT, closure duration, ratio, duration & VOT, VOT & ratio, and all three combined. The best Likelihood Ratio results were achieved using the combined /t/ & /k/ data in the VOT & ratio system. A study of three sibilants /s, z, ʃ/ focusing on both static features (intensity, CoG, SD, skewness, and kurtosis) and dynamic ones (CoG depending on F_2 vowel onset and offset) [3] revealed high speaker-specificity of CoG, SD, and intensity (while the latter highly depends on the recording circumstances).

The comparison of temporally distant speech samples, which is the topic of interest of the present study, is relatively understudied, although this is the typical setting in the context of forensic speaker comparisons. The extent of the temporal gap between recording sessions is different case by case, which might vary from a couple of months to several years (the longest documented interval is 27 years [6]).

The relatively long time lag between the first and the second recordings of the speech can be interpreted from the aspect of age-related speech changes (disregarding here other factors, like speech mode, health, sociophonetic variation, context, etc., that may affect the variation of speech parameters). However, age-related speech changes are not linear, they might be abrupt and are influenced by several factors, e.g., the age of the speaker at the time of recording.

With respect to the comparison of non-contemporary samples, Rhodes [2] studied 8 speakers (6 men and 2 women) between the ages of 21 and 49, in 7-year intervals. He reported large interspeaker variation in all parameters. F_1 changed between the analysed samples and decreased from 21 to 49 years of age (for all but one vowel). However, the change showed large individual differences. F_2 and F_3 showed similar but less generalizable tendencies within the 28 years but no clear tendencies were found in the 7-year interval comparisons. Consonants were compared both in contemporaneous and non-contemporaneous pairs of speech samples in [5]; however, the exact temporal distance between the latter ones is not known.

In the present study, we analyse speech samples from two age groups, young and middle aged speakers, i.e., the typical population of forensic speaker analysis, and compare samples that were recorded from them with a 10-year-long time lag.

Within the age interval under analysis, i.e. in 10 years, no large biological changes are expected due to aging, although minor differences resulting from lifestyle, work environment, smoking or hormonal changes (especially in women) might appear.

The realization of segmental parameters in part are language-dependent, which should be considered in the methodology of forensic analysis [1]. In our study, we analyse realizations of vowels and consonants of Hungarian language which differs from English in many respects [7].

The present explorative study raised the following questions. Is there in group-level difference in vowel or consonant realizations between speech samples recorded with a time lag of 10 years? How reliably can formant and spectral moment values be used in non-contemporary speaker comparison?

2. METHODS

The speech of eight male and eight female speakers was analysed. All of them participated in collection of speech samples for 'BEA' database [8], and approximately ten years later in another for "Longitudinal" database [9] (henceforward, 'LD'). LD was established in 2017, in order to investigate non-contemporary speech variability and its effect on speaker verification. All the recording circumstances of the two databases were kept identical. The speakers' age varied between 20 and 34 years at the first recording and between 30 and 44 years at the second one.

Two reading tasks (15 separate sentences, and a 13-sentence long text) were used from the two databases. 16 occurrences of five (interconsonantal) vowels /ɒ, aː, i, ɛ, eː/ and 8 (intervocalic) occurrences of five consonants /s, ʃ, p, t, k/ were selected. Target sounds were manually labelled and automatically measured in Praat [10]. In vowels, the first four formants' frequencies were measured at the midpoint of the duration with the Formant (burg) algorithm of Praat with the basic settings (5 formants in the range of 0−5000 Hz for men, and 0−5500 for women). For the consonants, the four spectral moments were measured in the sound files resampled at 22 kHz in a 0.01 s window at the midpoint of the time interval of the fricatives, and from the start of the release burst of the stops. In the case of alternative realizations of stops (spirantisation), the mid 0.01 s was selected for measurement. Outliers were eliminated by speaker, by vowel, and by recording session. The ratios of the median spectral values in the LD to the BEA were calculated ('LD to BEA ratio'). The inter-database, i.e. between-session differences were tested by linear mixed models (LMM) [11, 12, 13, 14, 15]. The dependent variable was one of the attested spectral features. After model selection (anova()), the fixed effects were RECORDING ('BEA' vs. 'LD', GENDER, and SPEECH SOUND allowing for their interaction, and the model included random slopes only on RECORDING by SPEAKERS. A simpler model was selected for kurtosis, where the factors CONSONANT and RECORDING were the fixed effects, and the model only included random intercepts by SPEAKER. *P*-values were obtained with Satterthwaite-approximation by anova().

We opted for LDA analysis [16], as in the context of speaker comparison, this method is used "to assess the speaker-specificity of a given variable and how useful it might be as a parameter in forensic casework" ([4]: 68). LDAs were run separately for the speech sounds and genders using all four spectral features measured for the specific speech sound. Spectral features were Z-transformed within the speech sound categories. A set of LDAs (2 speech sound groups * 5 speech sounds * 2 genders) was run within the first ('BEA') recording dividing the datasets into training and testing sets (.6 : .4 ratio for vowels, and .5 : .5 ratio for consonants for each speaker randomly). Another set of LDAs was trained on the dataset of the 'BEA' recordings, and tested on the dataset of the 'LD' recordings. The true positive and the true negative ratios were calculated over speakers. True positive ratio is the ratio of the items which were correctly classified to originate from the specific speaker to all the items that originate from the same speaker. True negative ratio is the ratio of the items correctly classified as not originating from a specific speaker to all the items that originate from another speaker.

3. RESULTS

3.1. Between-database comparison of spectral features

Median formant frequency values compared between the first ('BEA') and the second ('LD') recording showed large variability across vowels, speakers, and formants, and suggested no group-level changes over time in the studied time frame (Fig.1). F_1-data in women's pronunciation and the first two formants in men's pronunciation would suggest some possible change; however, these group-level differences did not appear on the speaker level, but in different speakers and vowels. No significant main effect of RECORDING for F_1 was found (LMM: Tab.1.) Although the interaction of the factors RECORDING and VOWEL was significant, post hoc tests revealed no significant F_1-shift tendency between recording sessions. RECORDING played a role neither as main effect, nor in any of the interactions on F_3. A general, small, significant increase was found in F_2 (estimated means: 1329 Hz, 1338 Hz, which is a negligible ratio of +0.006), and F_4 (estimated means: 3787 Hz, 3821 Hz, which is a negligible ratio of +0.009). However, post hoc tests for the significant threefold interaction did not reveal any vowel- and/or gender-specific tendency for any of these formants. This means that, no group-level change was detected between the two recording sessions for any of the formants in any of the vowels.

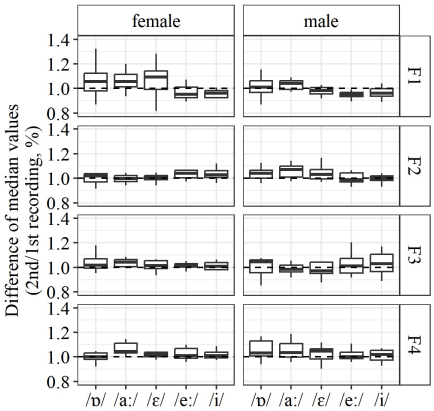

Figure 1: Change (%) of the formant values from the first ('BEA') to the second ('LD') recording (in 10 years) of the analysed vowels by gender.

factor	F	p	F	p
	F_1		F_2	
R	F(1,14.0)=1.8,	n.s.	F(1,14.3)=7.6	*
V	F(4,2251.3)= 1692.7	***	F(4,2143.0)= 1688.1	***
G	F(1,14.0)=22.6	***	F(1,14.03)=36.8	***
R:V	F(4,2251.3)=9.6	***	F(4,2143.3)=0.9	n.s.
R:G	F(1,14.0)=0.02	n.s.	F(1,14.3)=0.04	n.s.
V:G	F(4,2251.3)=14.8	***	F(4,2143.0)=9.7	***
R:V:G	F(4,2251.3)=1.3	n.s.	F(4,2143.3)=3.5	**.
	R_m^2=.693, R_c^2=.789		R_m^2=.756, R_c^2=.802	
	F_3		F_4	
R	F(1,14.0)=2.938,	n.s.	F(1,14.0)=15.7	*
V	F(4,2309.1)= 1136.3	***	F(4,2289.4)=14.3	***
G	F(1,14)=33.3	***	F(1,14.0)=5.7	*
R:V	F(4,2309.2)=0.5	n.s.	F(4,2289.4)=1.7	n.s.
R:G	F(1,14.0)=0.03	n.s.	F(1,14.0)=10.0	n.s.
V:G	F(4,2309.2)=0.9	n.s.	F(4,2289.4)=6.2	***
R:V:G	F(4,2309.2)=1.5	n.s.	F(4,2289.4)=3.5	**
	R_m^2=.284, R_c^2=.416		R_m^2=.104, R_c^2=.296	

Table 1: The results of the linear mixed models for the vowel formants. (R = recording, G = gender, V = vowel. Grey cells = significant differences.) (Sign. levels: 'n.s.' = not significant, '*' < .05, '**' < .01, '***' < .001).

Alternative realizations were found in some stops (non-burst release, but a fricative-like opening, i.e., with lower intensity, or spirantization/constant leak along the entire duration): in 10% of /p/, 2% of /t/, and 18% of /k/ sounds, altogether in the two databases. The 'LD to BEA' ratio of CoG showed larger interspeaker variability in the stops than in the sibilants (Fig.2). SD for women showed similar tendencies as CoG, but not for men. Between-session differences in skewness and kurtosis showed larger variability across speakers within the fricatives than CoG and SD. LMMs (Tab. 2) found a significant effect of RECORDING as main effect for skewness (estimated mean and change: +3.77, +2.53, which means a ratio of +1.67) and kurtosis (estimated mean and change: +28.21, +45.98, which means a ratio of +2.63). The interaction of the factors RECORDING and CONSONANT also significantly affected skewness and kurtosis. The post hoc test of these interactions found a significant difference in skewness of /p, k/ and kurtosis of /p, t, k/ between the two recording sessions. RECORDING as main factor did not have a significant effect on CoG and SD, while the interaction of RECORDING and CONSONANT was significant for these spectral moments, and the threefold interaction for CoG. However, the post hoc test of these interactions revealed no group-level tendencies in the change of CoG and SD between the two recording sessions.

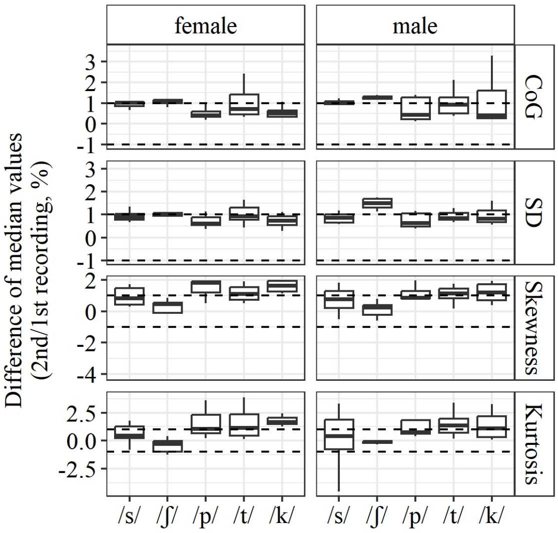

Figure 2: Median change of the spectral moments of the analysed consonants from 'BEA' to 'LD' (ratio).

3.2. Results of LDAs

The true negative ratios (TNR) did not considerably change between the two LDA scenarios (contemporary: 0.87–0.94, non-contemporary: 0.83–0.91). The true positive ratios (TPR) for the vowels decreased approximately to their halves (contemporary: 0.33−0.61, non-contemporary: 0.10−0.30), especially for men's production. The TPRs for the consonants were considerably lower than for the vowels (contemporary: 0.11−0.3, non-contemporary: 0.11−0.32). In the contemporary tests /ʃ/ for women (0.29), and /s/ for men (0.38), in the non-contemporary tests both sibilants for women (/ʃ/: 0.32, /s/: 0.24) and /ʃ/ for men (0.29) reached considerably higher TPRs than the further consonants which's TPRs stayed below 0.2.

factor	F	p	F	p
	CoG		SD	
R	F(1,14.1)=0.1	n.s.	F(1,14.0)=0.3	n.s.
G	F(1,13.9)=15.3	**	F(1,14.0)=0.8	n.s.
C	F(4,1186.2)= 1812.9	***	F(4,1209.1)=19.8	***
R:G	F(1,14.07)=2.6	n.s.	F(1,14.0)=0.8	n.s.
R:C	F(4,1186.0)=5.0	***	F(4,1209.2)=10.1	***
G:C	F(4,1186.2)=16.8	***	F(4,1209.1)=2.8	*
R:G:C	F(4,1186.0)=2.6	*	F(4,1209.2)=0.9	n.s.
	R_m^2=.844, R_c^2=.859		R_m^2=.089, R_c^2=.318	
	Skewness		Kurtosis	
R	F(1,14.3)=9.4	**	F(1,1228.2)=52.4	***
G	F(1,14.0)=0.73	n.s.		
C	F(4,1226.0)= 258.7	***	F(4,1228.0)=40.1	***
R:G	F(1,14.3)=0.1	n.s.		
R:C	F(4,1226.0)= 17.2	***	F(4,1228.1)=12.2	***
G:C	F(4,1226.0)= 2.0,	n.s.		
R:G:C	F(4,1226.0)= 0.9	n.s.		
	R_m^2=.444, R_c^2=.522		R_m^2=.163, R_c^2=.221	

Table 2: LMM results for the non-contemporary comparison of the spectral features (Sign. levels: 'n.s.' = not significant, '*' < .05, '**' < .01, '***' < .001).

4. DISCUSSION AND CONCLUSIONS

The present paper discussed the relevance of speaker-specific variability of eight spectral features', namely F_1−F_4 of five vowels and the four spectral moments of five voiceless obstruents between speech samples of Hungarian from a decade apart.

F_1 of the low, and mid-low vowels in women's pronunciation, and F_3, and F_4 in men's pronunciation showed larger interspeaker variability in the change between the two recordings. Only the F_2 and F_4 differences between the two recordings were found to be significant (LMM); however, their ratios of the estimated means were below 10%. As for the consonants, the CoG of the stops in men's pronunciation, and the skewness and kurtosis for most consonants showed large interspeaker variability in the comparison of the two recording sessions. Group-level change was found for the skewness of /p, k/, and the kurtosis of /p, t, k/. The low number and random occurrence of alternative realizations of stops do not explain the results of kurtosis and skewness.

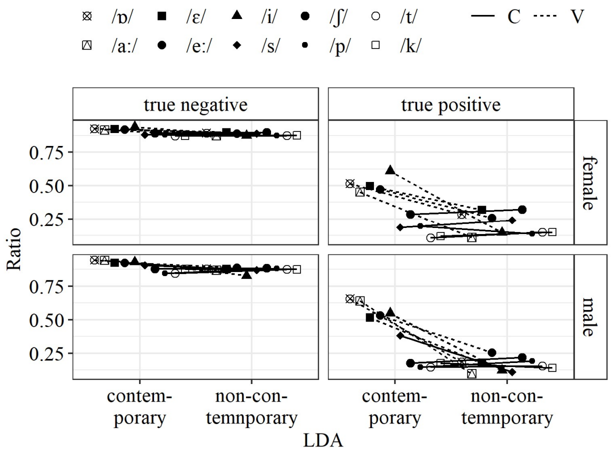

Figure 3: The true negative and the true positive values of LDAs within 'BEA' and between the two databases.

The true negative rates of LDA results did not vary considerably between the contemporary and non-contemporary testing. Its values ranged quite high, above 0.8 due to the relatively high number of speakers (at least for LDA). The TPRs for LDA tests using the spectral moments of the consonants were low, but they were higher for sibilants than for stops. These values did not change considerably between the two LDA scenarios. The TPRs for vowels were twice-trice higher in the contemporary than in the non-contemporary LDA tests. In the non-contemporary comparison, the TPRs for vowels and sibilants were similar.

Based on these results, we conclude that our hypotheses are not supported, as there is no considerable difference among the LDA tests using different speech sounds.

The limitations of our study are the relatively low number of speakers, and the low number of obstruents analysed. The results, however, suggest that the interspeaker differences in the between-recording variance in non-contemporary speech samples raise difficulties in forensic tasks.

5. ACKNOWLEDGEMENT

The research was funded by the National Research, Development, and Innovation Office (grant number: FK128814).

6. REFERENCES

[1] Rose, P. 2002. *Forensic Speaker Identification.* Taylor & Francis.

[2] Rhodes, R. W. 2012. *Assessing the strength of non-contemporaneous forensic speech evidence.* PhD thesis. The University of York.

[3] Fernandez, S. C. 2022. Speaker Specific Information in the Acoustic Characteristics of English Fricatives. *IJFL* 3 , 105–115.

[4] Kavanagh, C. M. 2012. *New consonantal acoustic parameters for forensic speaker comparison.* PhD thesis. The University of York.

[5] Earnshaw, K. 2016. Assessing the Discriminatory Power of /t/ and /k/ for Forensic Speaker Comparison using a Likelihood Ratio Approach. *IAFPA 25.* York, 92–93.

[6] French, J. P. F., Harrison, P., Windsor-Lewis, J. 2006. R v John Samuel Humble: The Yorkshire Ripper Hoaxer trial. *IJSLL* 13 , 256–273.

[7] Siptár, P., Törkenczy, M. 2007. *The Phonology of Hungarian.* Oxford University Press.

[8] Gósy, M. 2013. BEA – A multifunctional Hungarian spoken language database. *Phonetician* 105, 50−61.

[9] Gráczi, T. E., Huszár, A., Krepsz, V., Száraz, B., Damásdi, N., Markó, A. 2020. Longitudinális korpusz magyar felnőtt adatközlőkről. In Berend, G., Gosztolya, G., Vincze, V. (eds), *XVI. Magyar Számítógépes nyelvészeti konferencia. Proceedings.* University of Szeged. 103−114.

[10] Boersma, P., Weenink, D. 2022. *Praat: doing phonetics by computer.* http://www.praat.org/

[11] R Core Team 2022. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

[12] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *JSS* 67, 1−48.

[13] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *J SS* 82, 1−26.

[14] Lenth, R. 2022. emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.8.3, https://CRAN.R-project.org/package=emmeans.

[15] Bartoń, K. 2022. MuMIn: Multi-Model Inference. R package version 1.47.1, https://CRAN.R-project.org/package=MuMIn.

[16] Venables, W. N., Ripley, B. D. 2002. Modern Applied Statistics with S. Fourth Edition. Springer.