ARCHIVES OF ACOUSTICS
31, 4 (Supplement), 211–217 (2006)

CONFRONTING REPRESENTATIVE SPECTRAL STRUCTURES OF VOCAL TRACT PULSE RESPONSES

Z. WOJAN, W. LIS

Gdańsk University of Technology
Faculty of Electronics, Telecommunications and Informatics
Department of Marine Electronics Systems
Gabriela Narutowicza 11/12, 80-952 Gdańsk, Poland
e-mail: wojan@eti.pg.gda.pl

K. WOJAN

University of Gdańsk
Institute of East Slavonic Studies, Russian Language Chair
Wita Stwosza 55, 80-952 Gdańsk, Poland

(*received June 15, 2006; accepted September 30, 2006*)

Before the resources of a number of ethnic language systems can be confronted, multiple linguistic procedures must be applied to identify the possible common acoustic features in the utterances of the language users. There is a clear difference in the articulatory habits characteristic of the particular languages, which is why significant similarities must be sought in the oblique fragments of speech that convey the information code (sense) rather than in the acoustic representation of free speech. The paper discusses methods and results of automatic classification of selected lexemes of three language systems by confronting their digital representations. Digital representation includes sets of acoustic parameters as discussed in the previous OSA papers, which the authors termed as: *representative spectral structures of vocal tract pulse responses. Representative structures*... were produced by cepstral smoothing of averaged acoustic parameters taken from multiple utterances by speakers of different genders and ages. The paper includes spectrograms of the material used for confrontation and graphic illustrations of the results. Automatic classification of speech sounds of a pair of lexemes of two different languages is in fact the process of identifying the percentage proportions of convergent and divergent energies of lines in a set of parallel elements of both spectral matrixes being confronted, with time and frequency coordinates.

Key words: speech sound, information code, vocal tract, homomorphic analysis, homonymy.

1. Introduction

Confronting the parameters of speech sounds of two different languages, especially with the purpose to define the degree of similarity (or identity) between pairs of lexemes selected during linguistic confrontative analysis, is a risky business. The idea is not to compare pairs of words, i.e. inter-language homonyms [7]. This happens when

the acoustic form of a language 1 system lexeme is played back for language 2 users, who make clear links with an expression that they use in their own language practice and *vice versa* [4]. Speech sounds which belong to another system are associated with a specific element of a user's own language, because over the years the users have been "tuned" to identify the information code correlated with native ethnic speech articulation habits [2]. The brain recognises an alien lexeme as a "known" lexeme despite the "alien sound", frequently produced by the different phoneme structure of the other language system and also in isolation from the real semantic value of the expression just heard. The consequence may be a misunderstanding and the reasons for this are found in the term "false friends of interpreters". Intellectual verification of the language code contained in speech sounds is the perfect method, therefore inter-language homonyms do not require homomorphic analysis, but can be used as a model for acoustic similarity of lexemes. This similarity can be used as a basis for testing a digital analysis system and for determining the threshold for formal convergence.

The acoustic affinity between expressions of different languages may be the result of loan words or the diffusion of internationalisms or the fact that the languages have the same etymological roots and may be identifiable during playback by a researcher. To conduct a confrontative analysis of the resources of language pairs, linguists can formulate the criteria for classifying "acoustic similarity" based on the literature, their own specialist knowledge, research experience, etc. But the results carry traits of subjectivity, which is fundamentally against the overall principle of modern contrastive linguistics: a uniform research platform, compatible methodologies, standardised criteria for assessment enabling reliable comparisons of a wide range of languages concerning results obtained by scholars working in any far flung corner of the world [8, 9].

Seeking to identify inter-language homonym relations, inter-ethnic psycholinguistic backgrounds or trace areas of lexical common ground are research processes which force the researcher to conduct reliable, repeatable, objective assessments of acoustic community of expressions concerning the full range of speech sounds and its important fragments. To meet these complex and interdependent tasks which are both humanistic, philosophical and engineering and acoustic, there must be interdisciplinary teams of specialists processing the digital representations of speech signals obtained from samples of analogue utterances from ethnic language users representing different genders and ages. Below is an overview of the results produced by an interdisciplinary team [10–12].

2. Representative spectral transmittances of ethnic lexemes

Because of the redundancy of information, which is inherent to how speakers convey messages (content, sense), speech signals must be decomposed before they undergo inter-language confrontation. First to be able to use cepstral signal representations, the utterances must be subjected to homomorphic decomposition [5, 6]. This process helps to isolate sounds responsible for conveying intellectual content from the burden of speaker specific information about the parameters of the excitation signal

(larynx tone, noise). But the undesirable (in this case) live speech features cannot be separated easily or fully and cannot escape the deficiencies of digital processing. Not to be forgotten is the fact that the digital form of what is in fact an analogue speech signal is an imperfect approximation and the numerous mathematical operations that the digital representation undergoes introduce interference and false information, which increases the distance between the subject and the original model even more. Validity of mentioned comment appears to be heared when listening analog speach waveform after its digital processing. A realisation of these shortcomings is not enough, and it is important to remember that speech processing based on cepstral signal representations is the cause of a number of problems because of how difficult it is to interpret the results [1].

The transmittance of a linear system representing the vocal tract of a single speaker while he is articulating a specific utterance gives a good description of the system, but it cannot be used to discern universal features which shape the distinctive features of the acoustic layer of an ethnic word (lexeme). There are two fundamental factors which leave their negative mark on a single spectral transmittance: the individual structure of speech organs, and what is worse, the fact that the acoustic channel during articulation is not invariant to time and the times of operations which shape the excitation signal are enormous for confrontation purposes, although they are acceptable language rules. In an effort to identify features in speech signals, which the listener's brain would recognise as pure, intellectual information (sense), the authors of this paper developed a method for averaging the parameters of multiple cepstral representations to enable confrontation of transmittance of different ethnic lexemes. The results are "ethnic spectral structures" of pulse responses of a universal vocal tract, which are representative for the articulation of specific lexemes in a specific language system [3, 12].

The results had to be assessed for their correctness and usability. The method may lead to radical changes of results even for the smallest of parameter adjustments by the computational system, and the matter of cepstral structures is completely alien to human senses. This is why the semantic code (communicativeness) of speech was tested in ethnic multi-utterance structures of spectral transmittances by employing the method of public playback of a complete speech. In all instances the test utterance was obtained artificially as a result of a combination of an ethnically representative set of pulse responses of a universal vocal tract and a single excitation signal of any user of the language.

3. Automatic confrontation of representative spectral transmittances of ethnic lexemes

Figure 1 shows examples of spectral (representative) transmittance of Polish lexemes and spectrograms of the results of automatic confrontation of parallel elements of time and frequency matrices, which is what these structures in fact are. The system computes the modules from complex values of parallel elements of both matrices, divides and selects the values: those within the range $\{1/p,\ p\}$ (p – maximum value

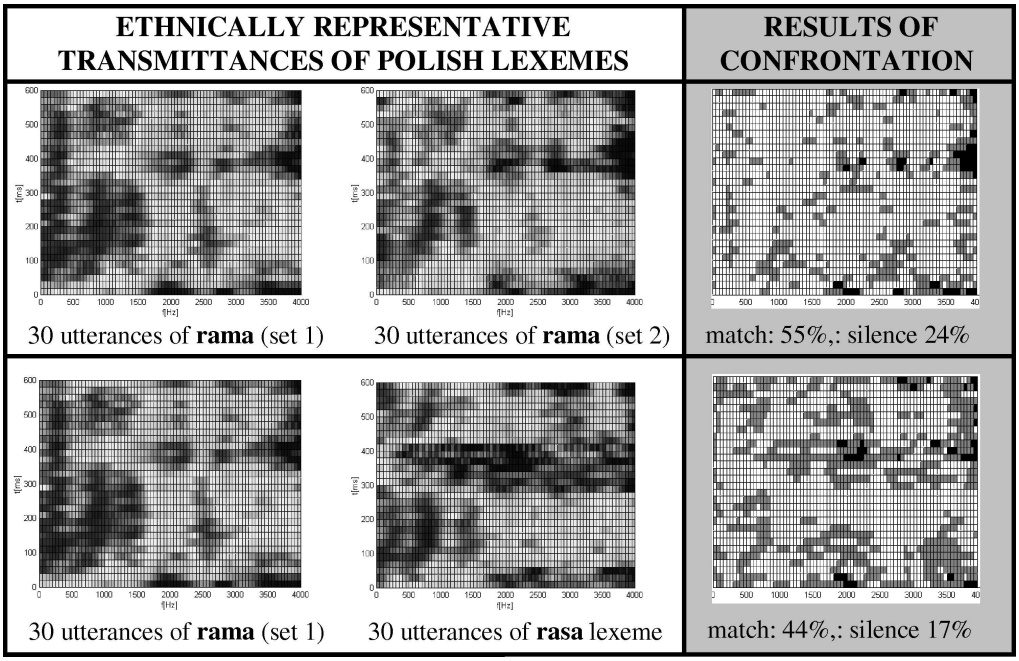

Fig. 1. Examples of automatic confrontation of ethnically representative spectral transmittances of Polish lexemes: **rama** (2 sets) and **rasa**.

Table 1.

Polish lexeme	Polish lexeme	$1/p \quad p$	Match [%]	Silence [%]	**Sum** [%]	Mismatch [%]
rama (set 1)	**rama** (set 2)	0.50–2.0	60.60	23.71	**84.31**	15.69
rama (set 1)	**rama** (set 2)	0.66–1.5	54.97	23.71	**78.68**	21.32
rama (set 1)	**rama** (set 2)	0.83–1.2	47.42	23.71	**71.13**	28.87
rama (set 1)	**rama** (set 2)	0.91–1.1	44.05	23.71	**67.76**	32.24
rama	**rana**	0.66–1.5	52.35	17.69	**70.04**	29.96
rama	**raja**	0.66–1.5	39.40	17.51	**56.91**	43.09
rama	**rasa**	0.66–1.5	44.56	16.84	**61.40**	38.60
rama	**rafa**	0.66–1.5	39.30	26.36	**65.66**	34.34
rana	**rafa**	0.66–1.5	40.30	20.34	**60.64**	39.36
rana	**rasa**	0.66–1.5	47.85	12.69	**60.54**	39.46
rana	**raja**	0.66–1.5	59.47	13.87	**73.34**	26.66
rafa	**raja**	0.66–1.5	43.91	21.40	**65.31**	34.69
rasa (Polish)	**paca** (Russian)	0.66–1.5	29.66	31.25	**60.91**	39.09
rana (Polish)	**paía** (Russian)	0.66–1.5	55.68	20.89	**76.57**	23.43
raja (Polish)	**raja** (Finnish)	0.66–1.5	43.73	15.55	**59.28**	40.72
aura (Finnish)	**paca** (Russian)	0.66–1.5	25.38	21.22	**46.60**	53.40

of product of parallel compared elements (arbitrary taken), in the examples quoted in the article $p = 1.5$) are classified as a "match", or "mismatch" if not found within the range.

We must not allow the decision-making system to classify those products which fit within the range when the results originate from very small value divisions (at noise level), this is why the system eliminates the elements in the matrices whose modules do not reach one per cent of the module's maximal value in each of the matrices (threshold). In an effort not to lose the proportions between the naturally quieter part of the utterance (match) and low values from cepstral analysis in high frequencies (highly accidental), the system analyses the correlation between the "sub-threshold values" and the duration of the utterance; when there is a match between parallel elements it is "silence"

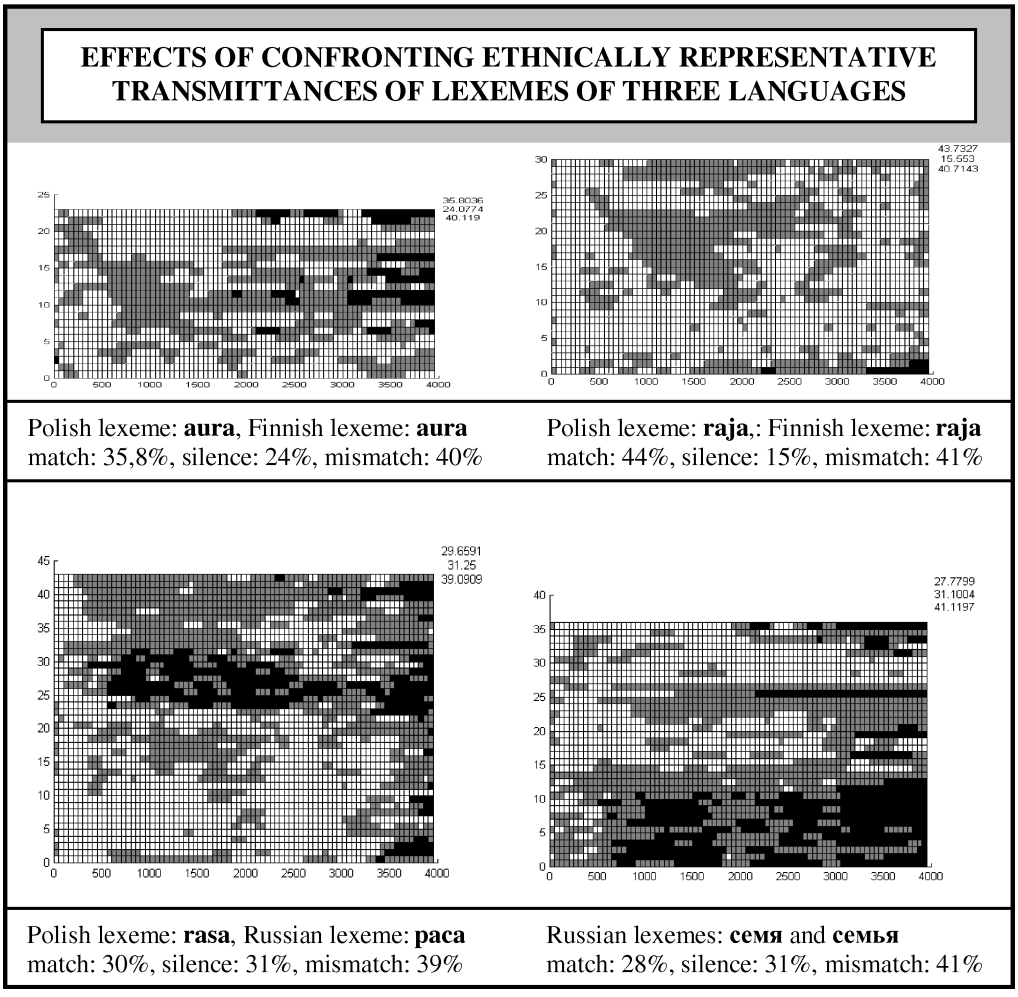

Fig. 2. Examples of automatic confrontation of spectral transmittances (ethnically representative) of pairs of lexemes in three languages. "Match" – light spots, "silence" – black spots, "mismatch" – grey spots.

and when the sub-threshold and over-threshold values come together it is classified as "mismatch".

Some results of automatic confrontation of ethnically representative spectral transmittances of pairs of lexemes in three languages are given in Table 1.

Figure 2 shows examples of the results of automatic confrontation. The selection was made based on linguistic criteria applications and a high degree of differentiation between Polish, Russian and Finnish languages and their lexemes. In the case of "**raja**" (Polish), "**raja**" (Finnish) and "**rasa**" (Polish) "**paca**" (Russian) lexemes full utterances were used (representative spectra of pulse response were selected of an ethnically universal vocal tract with an identical time axis range), for the pair of "**aura**" (Polish) and "**aura**" (Finnish) lexemes because articulation is completely different in Polish and Finnish, the duration of the utterance in Finish was limited to a value equivalent to the duration of an average equivalent utterance in Polish. It is interesting to observe the effect of confrontation of two lexemes (**семя** and **семья**) from a single language system (Russian), which are stressed differently.

4. Conclusion

At the present stage it is clear that automatic classification of lexeme confrontation to assess their acoustic similarity is ineffective. The percentage share of elements classified as identical within the overall number of elements in comparable matrixes of spectral transmittances is still too wide and there is concern about the overlap when phonetically similar lexemes are compared within a single language system. There are two possible sources of why the method has failed.

First, the forming of multi-utterance structures of spectral transmittances, in particular concerning the uniformity of time distribution of acoustic events within a set of summed up and averaged utterances; the averaging process needs further analysis.

Second, the concept of an algorithm for confronting complex values in parallel windows of time and spectral matrixes, choosing the right form for representing complex numbers derived from a direct comparison of values in the windows, the number, distribution and range of values both for cancelled confrontation and the range (threshold) to classify the parameters and for making the final decision "identical/not identical". The work will be continued.

References

[1] BASZTURA C., *Źródła, sygnały i obrazy akustyczne*, WKiŁ, Warszawa 1988.

[2] MOORE B. C. J., *Wprowadzenie do psychologii słyszenia*, PWN, Warszawa 1999.

[3] JASSEM W., *Podstawy fonetyki akustycznej*, PWN, Warszawa 1973.

[4] KURCZ I., *Psychologia języka i komunikacji*, Scholar, Warszawa 2000.

[5] SAPOŻKOW A., *Sygnał mowy w telekomunikacji i cybernetyce*, WNT, Warszawa 1966.

[6] TADEUSIEWICZ R., *Sygnał mowy*, WKiŁ, Warszawa 1988.

[7] WOJAN K., WOJAN Z., *Analiza akustyczna jako parametr klasyfikacji formalnej homonimii międzyjęzykowej*, XLIX Otwarte Seminarium z Akustyki OSA'2002, pp. 355–360, Warszawa-Stare Jabłonki, PTA Oddział Warszawski, 2002.

[8] WOJAN K., WOJAN Z., LIS W., *Akustyczny obraz słowa na tle mowy etnicznej*, XLIX Otwarte Seminarium z Akustyki OSA'2002, pp. 343–348, Warszawa-Stare Jabłonki, PTA Oddział Warszawski, 2002.

[9] WOJAN K., WOJAN Z., LIS W., *Digital analysis of speech signals as an instrument of comparative linguistics* [in Russian], Kaunas, 3, 54–66, 2003.

[10] WOJAN Z., WOJAN K., *Cyfrowa analiza mowy etnicznej – ekstrakcja kodu informacji*, 50. Otwarte Seminarium z Akustyki, pp. 395–399, Szczyrk-Gliwice, PTA Oddział Górnośląski, 2003.

[11] WOJAN Z., WOJAN K., *Selected problems of contrastive analysis of the acoustic images of lexemes of different languages* [in Russian], III Jornadas Andaluzas de Eslavística: lingüística, didáctica, folclorística, literatura y traducctión, Enrique F. Quero Gervilla, Ángela Salmerón Vílchez [Eds.], pp. 175–176, Granada 2004.

[12] WOJAN Z., WOJAN K., *Reprezentatywne dla danego języka spektra odpowiedzi impulsowej kanału głosowego*, 52. Otwarte Seminarium z Akustyki, pp. 191–194, Poznań-Wągrowiec, PTA Oddział Poznański, 2005.