Clustering analysis of phonetic and text feature vectors

Milan Jičínský
University of Pardubice
Faculty of Electrical Engineering and Informatics
Studentská 95, 530 02 Pardubice I, Czech Republic
milan.jicinsky@student.upce.cz

Jaroslav Marek
University of Pardubice
Faculty of Electrical Engineering and Informatics
Studentská 95, 530 02 Pardubice I, Czech Republic
jaroslav.marek@upce.cz

***Abstract*— Our goal is to show an example of using statistical methods to analyse some attributes of speeches. For this purpose, the New Year's Day speeches of Czech and Czechoslovak presidents are chosen. The aim of our study is researching similarities among these speeches and their recognizability through the history of Czechoslovak politics. All presidents are compared between each other. The comparison method is based on principal component analysis and cluster analysis. Important part is creating a feature vector. The feature vector doesn't have to be the same for successful clustering. There are many varieties and combinations of features that can be selected and used. Correlated variables must be discarded. The most significant features are chosen to represent and characterize the speaker. Some speakers can have something in common according to the chosen features. Or on the other hand they can differ much more from others. This kind of approach can help us to recognize a speech pattern of each spokesman independently.**

Keywords—clustering; New Year's Day speeches; President; feature vectors; voice analysis; energy; zero crossing rate; speech velocity; linguistics; phonetics; segmentation; frames; audio processing; speaker comparison; principal component analysis; cluster analysis

I. INTRODUCTION

Clustering is a very spread statistical method used in various field of research [1-3]. Well described fundamentals and clustering algorithms can be find in [4]. Applying this method even for audio files including a music [5] and voice recordings isn't an exception. Another example of using clustering in this case for segmentation and classification of audio files is well described in [6]. Comparing audio sets can be very efficient but it's important to know that only sampled recording or voice fragments cannot be a subject of clustering. It's because of representation of speech as a raw signal. This means e.g. millions of sampled values. Those values depend on time and content of speech. Those values have no use for us. It's necessary to make some audio processing first. Instead of using signal samples, each recording is represented by parameters also called features. These features can be put together to create a feature vector. Only the representation of recording as a feature vector is acceptable for further statistical analysis.

In this research, our attention has been given to find some similarities of New Year's Day speeches by using statistical methods. This has been chosen example of showing how to extract information from available data and using them for researching speaker similarities considering the text and speech itself. The speech of Czech and Czechoslovak presidents can be characterized by various voice characteristics such as zero crossing rate, log energy, speech velocity, spectral energy and many more. To reduce the high-dimensional data, the principal component analysis and the hierarchical cluster analysis will be used. We will take a look at possibilities of using statistical method, phonetic analysis and mathematical linguistics for comparing of political speeches. The main goal is to show how recordings can be analysed and compared among each other in different way than we are used to. Different scientific approach can be achieved just by linking of linguistics, phonetics and statistical methods using clustering algorithms.

By help of linguistic characteristics and voice characteristics of speakers the similarities of New Year's Day Speeches of Czech presidents can be measured. The next step is to explore the differences between the result clusters. These results can show us the partial influences of speaker's characteristics. Our aim is to obtain the best results using and combining only other phonetic and text based features.

II. SAMPLE OF SPEECHES

A. Data

The data reserved for our research come from [7]. This is web audio archive containing almost every single presidential speech since 1935. The sources of these speeches and their transcriptions are archive of president's office, linguist Jaroslav David and Moravian Library. But thanks to Český rozhlas everything is in one place.

Each speech was recorded separately and sampled at 16 kHz using Audacity software. As mentioned before the recording must be edited. Otherwise it couldn't be used in our case. First, it is necessary to get rid of all parts containing a music, long silence or even a voice of moderator who has an introduction speech at the very beginning of recording. Finally, the data are ready to be examined.

B. *Measuring of voice parameters*

Since we have the recordings edited every speech should be segmented into smaller parts called frames. The frames are typically 20ms long and they are overlapping each other right in the half. Segmentation is followed by parameterization step. During the parameterization, the features are evaluated for each frame. These features can be divided into some groups. We can distinguish basic, spectral and cepstral features. Each of those groups can be considered either static or dynamic. Static features are computed exactly per the corresponding formulas. They must be calculated before the dynamic features. It's because dynamic features are given by static ones. While the static parameters of feature vector have their own meaning in sense of signal analysis and they are related to frequency and other measurable parameters, the dynamic features only express changes of the static values among frames. In the various publications, the dynamic features are also called delta features because it can be considered as the mathematical derivative of original parameter. Expression "delta" is used for the first derivatives. The second derivatives are called "delta-delta". The example of measured values for chosen speech is given in Table I. These numbers are the result of analysis for Masaryk's speech dated 1935 (left column) and Havel's speech dated 1996 (right column).

TABLE I. AN EXAMPLE OF SPEECH FEATURE VALUES

Feature type		**Speech feature vector**		
		Feature name	***Values***	
Static	***Basic***	Speech velocity	1.335	1.961
		E – energy	9.333	8.573
	Spectral	Bk – bin 0 – 500 Hz	11.283	10.257
		Bk – bin 500 – 1000 Hz	9.263	8.946
		Bk – bin 1 – 1,5 kHz	8.816	8.812
		Bk – bin 1,5 – 2 kHz	8.138	8.416
		Bk – bin 2 – 2,5 kHz	7.947	8.009
		Bk – bin 2,5 – 3 kHz	8.056	7.837
		Bk – bin 3 – 3,5 kHz	8.220	7.832
		Bk – bin 3,5 – 4 kHz	8.143	7.315
Dynamic (delta)	***Basic***	|ΔE| – delta energy	0.245	0.363
		|ΔΔE| – delta-delta energy	0.346	0.461
	Spectral	|ΔBk| – bin 0 – 500 Hz	0.294	0.413
		|ΔBk| – bin 500 – 1000 Hz	0.290	0.475
		|ΔBk| – bin 1 – 1,5 kHz	0.255	0.492
		|ΔBk| – bin 1,5 – 2 kHz	0.214	0.492
		|ΔBk| – bin 2 – 2,5 kHz	0.161	0.467
		|ΔBk| – bin 2,5 – 3 kHz	0.162	0.465
		|ΔBk| – bin 3 – 3,5 kHz	0.171	0.479
		|ΔBk| – bin 3,5 – 4 kHz	0.200	0.479

[a.] Source: own.

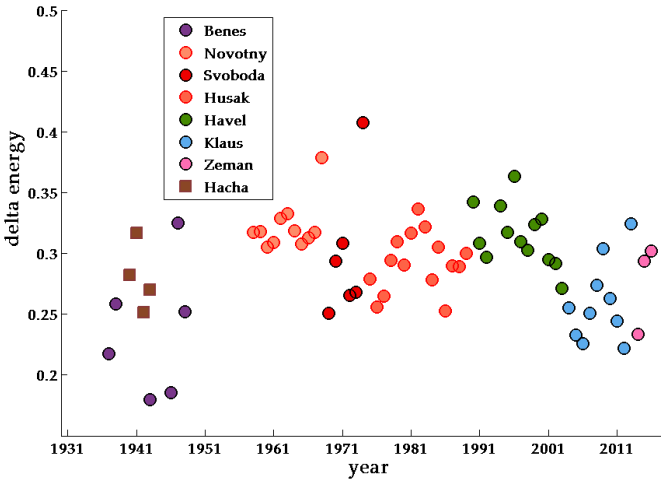

Fig. 1. Delta energy representing voice parameters. Source: own.

While energy and speech velocity were shown in [8], first and second derivative of energy values are given by Fig. 1 and Fig. 2. As can be seen a range of chosen feature values is completely different. Speech velocity depends on pace of the speaker. The number shows how many words were said within one second. Typical range is from 0,9 to 2,5. Average energy ranges between 7 and 10. As for spectral energy *Bk*, the situation is very similar but it differs depending on which bin was used. On the other hand, however delta features ordinarily have very low values. This is caused by the fact that the differences between frames are very small positive or negative numbers and they are even approaching to the zero value. This is the reason why original delta features were not used. Because their mean value is almost equal to zero. Instead of it the mean value was calculated from absolute value of those features. This is the only way how to use these dynamic parameters for clustering and it can be considered as a slightly different scientific approach. Finally, we decided not to include a Zero crossing rate due to the very high variance of values within one president. This variance is also shown in [8].

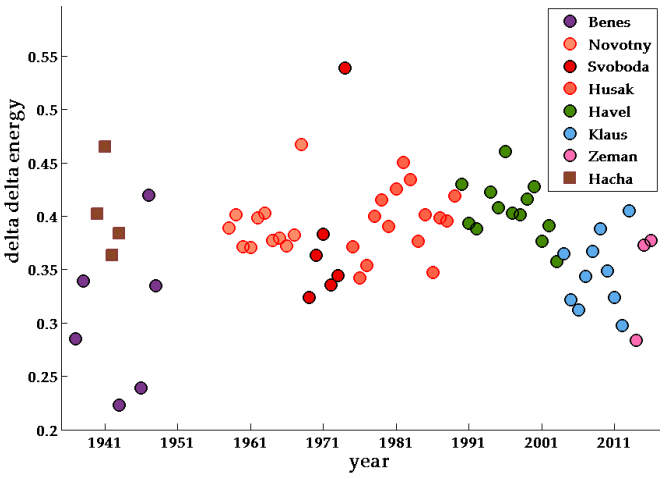

Fig. 2. Delta delta energy representing voice parameters. Source: own.

C. *Text feature extraction*

Some text related features can be defined and used as feature vector the same way as the phonetic parameters. Special feature vector representing some text characteristics were designed for the purposes of the article. Some of those features were already discussed in [8]. Total number of words has a relation with length of the whole speech. An amount of different words is the result of saying the same thing in different ways or just simply not intending to repeat the same words. Even sort of language richness can be related to this parameter. Even length of words plays an important role. While conjunctions and prepositions doesn't vary too much, the most contained word can easily characterize the speaker. Table II is an example of text feature vector. It's organized the same way as previous table. It means that the left column is reserved for Masaryk and the right one for Havel. The dates of chosen speeches remain the same.

All speech processing has already been done in [8]. Even more details about previous steps can be find there too. Fundamentals of audio processing, segmentation and parameterization are included. Unfortunately, only basic static features were used and described (energy, speech velocity, zero crossing rate). The rest of features will be named and shown in upcoming yet unpublished paper [9].

TABLE II. AN EXAMPLE OF TEXT FEATURE VALUES

Feature type		**Text based feature vector**		
		Feature name	***Values***	
TEXT	***Numbers***	Total number of used words	259	2749
		Amount of different words	190	1390
		Mean length of words	5.514	5.206
		Most dominant length of words	7	2
	Words	The most contained word	nation	country
		Most used conjunctions and prepositions	and	and

b. Source: own.

III. THE CLUSTERING

A. *The Principal Component Analysis*

Some correlation between the voice characteristics of the speeches occurred, and that's why it is better to use principal components analysis.

In the Principal Components Analysis (PCA), the data are summarized as a linear combination of an orthonormal set of the vectors. The first principal component accounts for as much of the variability in the data as possible, and each successive component represents as much of the remaining variability as possible. This is the same as performing the singular value decomposition of the covariance matrix var **X** = **U D V,** where **D** is diagonal matrix of eigen-values, and **U**, **V** are orthonormal. Cf. [1].

The results of PCA are shown in Fig. 3-7. We use up to three principal components.

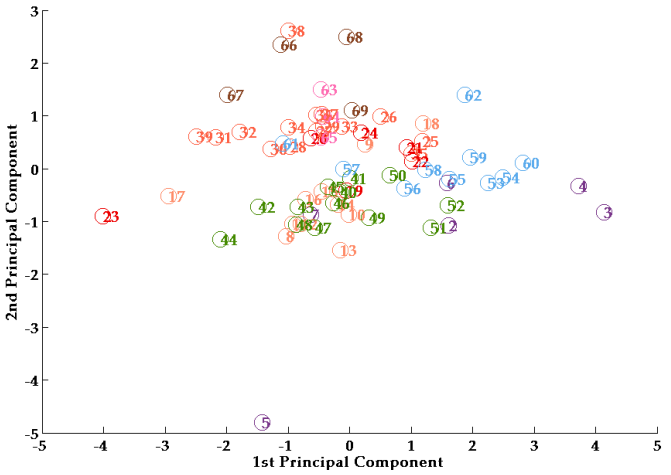

Fig. 3. PCA using 4 speech features. Source: own.

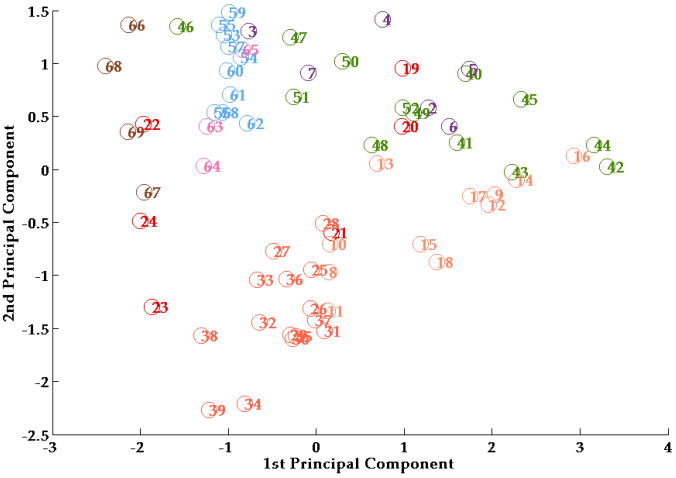

Fig. 4. PCA using 3 text features. Source: own.

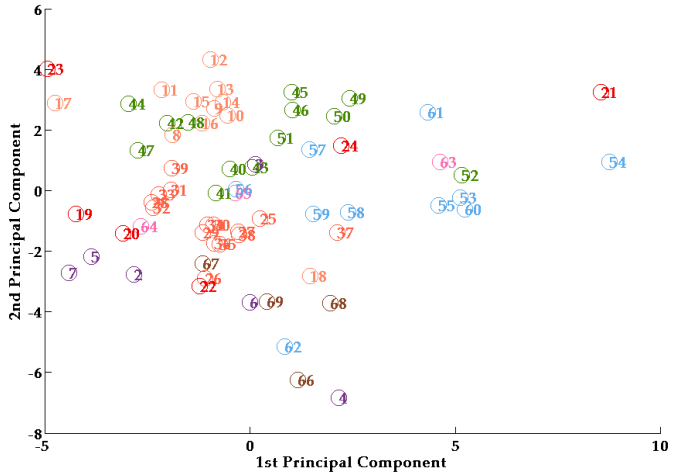

Fig. 5. PCA using 20 speech features. Source: own.

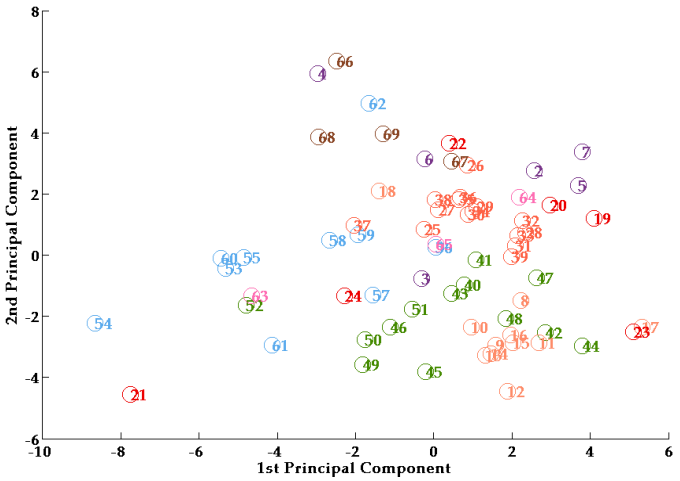

Fig. 6. PCA using 4 speech and 3 text features. Source: own.

It is necessary to know that results of the clustering may differ depending on chosen feature vector. That is why designing the vector can be considered as the most important part. More combinations were tried during our research. The partial component analysis was realized firstly for four parameters representing only speech (voice) - average energy, speech velocity, delta energy and delta-delta energy. The results can be seen at Fig. 3. Then we came up with idea of extracting only text based features. So, the second experiment were made using three representative linguistic parameters - total number of words, amount of different words, average length of words. Graphical interpretation is at Fig. 4. Then we tried to use the same feature vector as defined in Table I. Results of PCA for this vector containing twenty parameters are presented at Fig. 5. After these three tries we wanted to combine both linguistic and phonetic characteristics. And so, the next feature vector contains original twenty values of voice parameters plus three text based. This is shown in Fig. 7. The feature vector containing the fewest number of parameters but still combining text and voice parameters is used for PCA at Fig. 6.

The proportions of the principal component on the total variability of the original data for the different features are shown in the Table III.

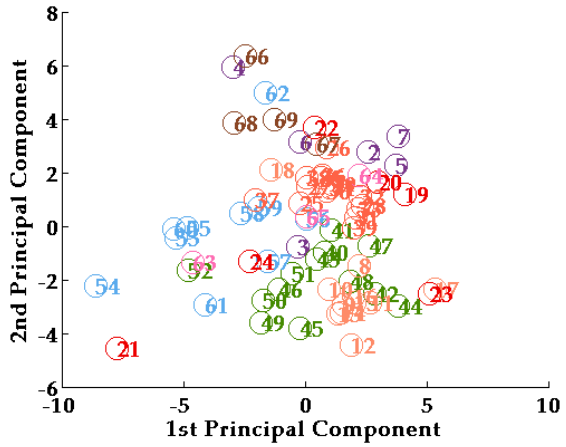

Fig. 7. PCA using 20 speech features and 3 text features. Source: own.

TABLE III. RESULTS OF PCA

Total variability	First component	First and second component	First, second and third component
Data20+0	55.6 %		
Data0+3	66.5 %	99.7 %	
Data4+0	40.3 %	71.1 %	82.3 %
Data4+3	35.5 %	63.4 %	75.2 %

[c.] Source: own.

We can see that the largest value of proportion on total variability occurs for the features data0+3. Thus, it is clear, that the author of the text is better defined by the text characteristics than the speaker described by the speech characteristics.

However, we will use the data4 file for further clustering, which is less suitable for criterion of total variability proportion maximization. The reason why the data 4+3 was used is an effort to work with both types of characteristics.

By help of another neighbour algorithm and the principal components we tried to find the clusters of chosen speeches.

The results obtained by using the hierarchical clustering are given in Table IV.

B. *Cluster Analysis*

In the table, we calculate the measure of membership into clusters for all speeches according to formula

$$\mu\left(x_i, v_j\right) = 1 - \frac{d\left(x_i, v_j\right)}{\max d\left(x_i, v_r\right)}, i = 1, \ldots, n; r = 1, \ldots, R, \qquad (1)$$

where d is the Euclidian distance, R is the number of patter objects. The most significant measure of membership is bold.

As for the Table IV, C1, ..., C8 are designations for patterns defined as centre of gravity of the first three components of PCA realized for 8 presidents and obtained for data containing 4 voice features and 3 text based. There are only 8 pattern categories because we have 8 presidents. Masaryk, Gottwald and Zápotocký are removed from further statistical analysis due to the lack of data. Each of them has only one short recording available. Discarding them makes the PCA results better. Each president should fit into the category with the same number as his ranking in the table. That means Beneš belongs to C1, Novotný to C2 and so on.

The only one whose classification is 100% accurate is the president of protectorate Hácha. He had significantly lower energy during his speeches. Havel has the only one wrongly classified speech within year 2003. Beneš and Novotný are fitting the right patterns too. As for those presidents, we can be satisfied with the final score. Results provide us an information that these presidents were very specific and they can be easily recognized and separated from the rest. They wanted their speeches to be the reflections of their own opinions.

TABLE IV. RESULTS OF CLUSTERING

president	C1	C2	C3	C4	C5	C6	C7	C8
Beneš 37	**26%**	9%	16%	12%	14%	11%	12%	0%
Beneš 38	**29%**	9%	17%	17%	10%	0%	12%	7%
Beneš 43	**20%**	16%	14%	10%	18%	13%	9%	0%
Beneš 46	**27%**	0%	12%	12%	4%	12%	17%	16%
Beneš 47	18%	12%	18%	**20%**	11%	0%	13%	9%
Beneš 48	**25%**	0%	15%	16%	5%	6%	19%	14%
Novotný 58	18%	13%	17%	**19%**	11%	0%	12%	10%
Novotný 59	10%	**26%**	16%	13%	22%	5%	8%	0%
Novotný 60	7%	**22%**	17%	13%	24%	7%	9%	0%
Novotný 61	7%	**22%**	17%	13%	25%	8%	9%	0%
Novotný 62	9%	**27%**	15%	13%	22%	6%	8%	0%
Novotný 63	7%	**24%**	15%	12%	23%	9%	8%	0%
Novotný 64	10%	**24%**	15%	12%	22%	9%	8%	0%
Novotný 65	8%	**24%**	16%	12%	24%	8%	8%	0%
Novotný 66	10%	**25%**	15%	12%	22%	8%	8%	0%
Novotný 67	9%	**25%**	16%	13%	23%	7%	8%	0%
Novotný 68	8%	22%	17%	13%	**24%**	7%	9%	0%
Svoboda 69	21%	0%	15%	14%	6%	12%	**20%**	11%
Svoboda 70	19%	19%	18%	17%	16%	0%	9%	3%
Svoboda 71	16%	13%	19%	20%	13%	0%	12%	8%
Svoboda 72	0%	4%	12%	9%	14%	**36%**	17%	8%
Svoboda 73	19%	0%	18%	22%	4%	0%	20%	18%
Svoboda 74	3%	**26%**	19%	19%	22%	0%	9%	2%
Husák 75	2%	10%	18%	14%	19%	**21%**	15%	0%
Husák 76	14%	3%	**27%**	23%	11%	3%	20%	0%
Husák 77	18%	2%	19%	**22%**	5%	0%	19%	15%
Husák 78	6%	2%	24%	**30%**	8%	0%	21%	9%
Husák 79	12%	7%	22%	**25%**	10%	0%	16%	8%
Husák 80	6%	7%	22%	**28%**	10%	0%	17%	10%
Husák 81	5%	13%	22%	**24%**	14%	0%	14%	8%
Husák 82	7%	12%	21%	**24%**	13%	0%	14%	9%
Husák 83	10%	15%	21%	**22%**	15%	0%	12%	4%
Husák 84	8%	7%	22%	**28%**	10%	0%	17%	9%
Husák 85	19%	4%	22%	**22%**	9%	0%	17%	6%
Husák 86	10%	4%	21%	**27%**	8%	0%	18%	11%
Husák 87	5%	0%	16%	17%	7%	14%	**25%**	17%
Husák 88	2%	0%	20%	26%	5%	1%	**24%**	22%
Husák 89	5%	21%	**22%**	21%	21%	0%	10%	0%
Havel 90	10%	18%	19%	15%	**22%**	6%	10%	0%
Havel 91	14%	19%	18%	15%	**20%**	5%	9%	0%
Havel 92	8%	27%	16%	14%	**22%**	5%	8%	0%
Havel 94	10%	19%	17%	13%	**23%**	8%	9%	0%
Havel 95	9%	27%	16%	14%	**22%**	5%	8%	0%
Havel 96	4%	16%	17%	13%	**22%**	16%	12%	0%
Havel 97	6%	17%	17%	12%	**23%**	15%	11%	0%
Havel 98	11%	24%	18%	15%	**21%**	3%	8%	0%
Havel 99	9%	25%	16%	13%	**23%**	6%	8%	0%
Havel 00	7%	15%	15%	11%	**21%**	20%	11%	0%
Havel 01	6%	13%	16%	11%	**21%**	20%	12%	0%
Havel 02	12%	17%	16%	11%	**21%**	13%	10%	0%
Havel 03	6%	7%	22%	**28%**	10%	0%	17%	10%
Klaus 04	5%	13%	22%	**24%**	14%	0%	14%	8%
Klaus 05	7%	12%	21%	**24%**	13%	0%	14%	9%
Klaus 06	10%	15%	21%	**22%**	15%	0%	12%	4%
Klaus 07	8%	7%	22%	**28%**	10%	0%	17%	9%
Klaus 08	19%	4%	22%	**22%**	9%	0%	17%	6%
Klaus 09	10%	4%	21%	**27%**	8%	0%	18%	11%
Klaus 10	5%	0%	16%	17%	7%	14%	**25%**	17%
Klaus 11	2%	0%	20%	**26%**	5%	1%	24%	22%
Klaus 12	5%	21%	**22%**	21%	21%	0%	10%	0%
Klaus 13	10%	18%	19%	15%	22%	6%	10%	0%
Zeman 14	0%	8%	15%	12%	16%	**27%**	16%	6%
Zeman 15	18%	0%	13%	15%	4%	9%	20%	**22%**
Zeman 16	0%	2%	14%	13%	11%	**28%**	20%	12%
Zeman 17	12%	10%	20%	**23%**	11%	0%	14%	8%
Hácha 40	11%	0%	12%	15%	4%	9%	20%	**29%**
Hácha 41	5%	0%	18%	24%	4%	0%	22%	**26%**
Hácha 42	9%	0%	12%	15%	4%	12%	21%	**26%**
Hácha 43	12%	0%	14%	16%	4%	8%	21%	**24%**

[d.] Source: own.

On the other hand, Svoboda, Klaus and Zeman are hard to distinguish. They have zero positive hits according to Table IV. Absolutely the worst classification can be recognized at speeches of president Klaus. His speeches are the most like president Husák.

IV. ALTERNATIVE POSSIBLE SOLUTION

Commonly used methods of comparing speeches are basically based on cepstral features and Hidden Markov Models. The clustering can be made using Mel-Frequency Cepstrum Coefficients (MFCC) and Linear Prediction Cepstral Coefficients (LPCC). Alternatively, even more efficient methods (especially for classification of speaker) use Gaussian Mixture Models, artificial intelligence (neural networks) and the most recent method called i-vectors. As for recognition of speaker it is very popular to create robust and text-independent recognition systems nowadays. Dynamic Time Warping (DTW) and HMM can't be used for this purpose. Our intention was to compare president speeches not the biometric recognition and classification of speaker. This is the reason why an advanced techniques and methods were intentionally omitted. More information about using HMM provides [10]. So, the alternative way can be realized by researching only recordings and omitting the transcript analysis.

In this case the gist of article is to search for similarities in president speeches according to the recordings and transcripts. Feature vectors were created for this purpose as noted above in previous chapter. These feature vectors combine the most important characteristics of speeches and text. They can be used for separated clustering or for linking text and speech together. Many more combinations of features may exist. This leads to the opportunity of further research in this field of study.

V. CONCLUDING REMARKS

Unfortunately, the dendrogram isn't suitable for graphical interpretation of clustering due to the number of president speeches. Using the feature vector containing twenty features as shown above doesn't bring expected results. This is influenced by correlation among most of these parameters. Even reducing the number of features to only four (average energy, speech velocity, delta energy and delta-delta energy) doesn't help that much to detect the speaker. The combination of text features and most significant phonetic features leads to the best results. Adding the cepstral features would be probably the best option for improving the results and then it could end up better. But thanks to the Table IV, the probabilities of belonging to the right cluster are relatively high. It is still very efficient even if some speeches of different presidents were clustered together. According to that fact we can find presidents with low rate of individualism expressed in the speeches – Zeman, Klaus and Svoboda. The rest can be more easily distinguished by their attributes that differ from one another. This research also proves that cepstral features are very hard to replace. Using cepstral features, Hidden Markov Models, Gaussian Mixture Models and neural networks may still provide better results, but doesn't allow to combine these features the way we did or they are not using any features at all.

ACKNOWLEDGMENT

This research was supported by the Internal Grant Agency of University of Pardubice, the project SGS 2017 024.

REFERENCES

[1] Steinbach, Michael, et al. "A comparison of document clustering techniques." In: KDD workshop on text mining. 2000. p. 525-526.

[2] Ruspini, Enrique H. "Numerical methods for fuzzy clustering." *Information Sciences*, 1970, 2.3: 319-350.

[3] Vesanto, Juha; Alhoniemi, Esa. "Clustering of the self-organizing map." IEEE Transactions on neural networks, May 2000, 11.3: 586-600.

[4] Jain, Anil K.; Dubes, Richard C. "Algorithms for clustering data." Prentice-Hall, Inc., 1988.

[5] Levy, Mark; Sandler, Mark. "Structural segmentation of musical audio by constrained clustering." IEEE Transactions on Audio, Speech, and Language Processing, January 2008, 16.2: 318-326.

[6] Lu, Lie; Zhang, Hong-Jiang; Jiang, Hao. "Content analysis for audio classification and segmentation." IEEE Transactions on speech and audio processing, December 2002, 10.7: 504-516.

[7] URL: http://www.rozhlas.cz/zpravy/data/_zprava/od-tgm-k-zemanovi-poslechnete-si-vanocni-a-novorocni-projevy-vsech-prezidentu--1436738. [accessed. 2017-08-13].

[8] Jičínský, Milan; Marek, Jaroslav. "New Year's Day speeches of Czech presidents: phonetic analysis and text analysis." In: IFIP International Conference on Computer Information Systems and Industrial Management. Springer, Cham, May 2017. p. 110-121.

[9] Jičínský, Milan, "Features reserved for clustering of New Year's Day speeches of Czech presidents," unpublished.

[10] Abdallah, Sayed Jaafer; Osman, Izzeldin Mohamed; Mustafa, Mohamed Elhafiz. "Text-independent speaker identification using hidden Markov model." World of Computer Science and Information Technology Journal, 2012, 2.6: 203-208.