Multimedia Tools and Applications (2019) 78:15575–15588
https://doi.org/10.1007/s11042-018-6933-1

Language ranking based on frequency varieties of phones

Mariusz Ziółko[1] · **Stanisław Kacprzak**[1]

Received: 26 March 2018 / Revised: 24 October 2018 / Accepted: 22 November 2018 /
Published online: 6 December 2018

Abstract
Phones for 239 non-annotated languages were selected by automatic segmentation based on changes of energy in the time-frequency representation of speech signals. Phone boundaries were set at location of relatively major changes in energy distribution between seven frequency bands. A vector of average energies calculated for eleven frequency bands was chosen as the representation of a single phone. We focus our research on an unsupervised comparison of phone distribution in 239 languages. Using the hierarchical clustering method, the relationship between the number of clusters and Ward's distance was determined. A mathematical model is proposed to describe this dependency. Its four parameters are determined for each language individually to model the relationship between the number of clusters and the frequency diversity of phones contained in clusters. We used these relationships to compare languages and to create their ranking based on the size of phone varieties in the frequency domain.

Keywords Speech technology · Frequency analysis · Language ranking

1 Introduction

A phone is a sound of speech, and a phoneme is an abstract representation of a phone. The main difference is that phones are characterized by physical features, such as the distribution of energy in frequency bands, while phonemes have a linguistic descriptions as speech elements. In other words, phones can be extracted from speech by electronic devices while phonemes are distinguished by the sense of hearing supported by linguistic knowledge. Phones and phonemes are not uniquely assigned to one another. From the physical point of view, speech signals are strongly distorted by the speaker's individual characteristics such as: sex, age, intonation, and emotional state. Additional distortions result from

✉ Mariusz Ziółko
ziolko@agh.edu.pl

Stanisław Kacprzak
skacprza@agh.edu.pl

[1] Faculty of Computer Science, Electronics and Telecommunications, AGH University of Science and Technology, al. Mickiewicza 30, 30-059 Kraków, Poland

co-articulation effects as a significant impact of neighbouring phones. All these phenomena strongly affect the physical properties of phone articulation. It is therefore reasonable to ask how, in spite of these distortions, speech signals are accurately analyzed by the human brain, and how electronic devices can be improved to enhance voice communication between humans and computers.

Quentin D. Atkinson in his article [2] suggested that the founder effect may operate on human languages. It means that expansion should progressively reduce phonemic diversity with increasing distance from the point of origin. His model points to central and southern Africa as the location of where the first languages may originate. Atkinson examined geographic variation in phonemes using data taken from 504 languages described in the World Atlas of Language Structures (WALS). His article [2] provoked immediate criticism (e.g. [22]) and had numerous citations. His opponents suggest that taking into account historical processes like migrations, conquests, and borrowings would explain language evolution more credibly than the founder effect solely. The scientific controversy about Atkinson's hypothesis motivated us to conduct an independent study to assess his suggestions. Unlike Atkinson [2], our approach was based on analyzing phones instead of phonemes.

Every language articulation exploits only a small part of innate human abilities. Young children are able to learn a spectrum of sounds broader than those existing in any particular language. Their individual articulation abilities are shaped by the culture that motivates children to master some phones and lose the ability to produce others at the same time.

We compared phones for 239 languages spoken by about 96% of the world's population. Our approach is an unsupervised speech research study. Unsupervised methods make it possible to analyze any language without prior linguistic knowledge. These methods try to mimic the way in which human sense analyze speech and infants learn language by simply being exposed to it.

Most approaches to automatic partition of speech into separate units do it in two steps [14]: segmentation joined with parametrization, followed by clustering. We also used this approach.

Frequency analysis is the first step in speech processing, by people and usually also by electronic devices. Computer analysis makes it possible to partition speech signals into segments which are characterized by relatively constant energy distribution in the frequency domain. Additionally, more precise frequency analysis makes it possible to parametrize phones. Next, these parametrization were used to determine the probability density of phone distribution in the 11-dimensional frequency domain.

Cluster analysis aims to reveal similarities between related phones collected in a data set [10]. Groups of similar elements (frequently associated with different variations of the same speech segment) were extracted by the clusterization of phones.

For any language the number of clusters which group similar phones and acceptable phone deviations in the frequency domain is not precisely defined. These two quantities, however, are closely tied. The greater the value of permissible deviations within the clusters, the smaller is the number of clusters identified with different representations of hypothetical phone representatives. The main goal of our publication is the presentation of experimentally determined dependencies of the number of phone representatives from permissible changes inside the clusters that group similar acoustic elements. The nature of this dependence is the same for all languages, although some characteristic differences are visible. We used these deviations as the basis for ranking languages.

Frequency analysis of phones allows us to calculate spectral properties in order to compare world languages. Such analysis provides information about languages from an articulation point of view [11, 18]. It is natural to expect different pronunciations between

different languages. Computer analysis uses signal processing methods to find the frequency properties of speech. Phone comparison between languages brings new and sometimes unexpected conclusions. Precise analysis of multi-linguistic speech aims to provide answers to the following question: how different are phones used in different languages and what are the individual features which characterize phone distributions?

This paper consists of five main parts. Chapter 2 introduces the database used to analyze the languages of the world. The next two chapters briefly describe the frequency method for automatic extraction of phones and their parametrization. Chapter 5 provides the basics of the clusterization method. The most important part of the paper is presented in Chapter 6, where we propose two methods for language characterization. They based on the dependence of the number of clusters from the Ward's distance obtained during the hierarchical clustering. Chapter 7 presents the results of calculations and suggestions how they can be interpreted. Chapter 8 concludes the paper.

2 Data acquisition

Vast volumes of speech recordings are not transcribed and do not have time annotations. Adding such annotations is an expensive and time-consuming process. Our motivation to develop a universal method for automatic extraction of phones form non-annotated speech is a need to compare phones of vast number of languages which do not have transcribed training data corpora. Therefore, fully automatic segmentation followed by phone analysis is extremely useful.

The diversity of languages can be verified by a computer analysis of speech recordings. To analyze frequency features of languages it is necessary to collect speech samples for hundreds of languages. Gathering speech recordings of appropriate quality and length is not an easy task. Results of analysis can be relevant if the duration of recordings for each language are sufficiently long. We have not found a database with speech recordings, created for scientific research and containing several hundred languages of the world. The Global Recordings Network (GRN) website [6] is a source of vast volumes of language recordings. GRN is a provider of Bible audio materials in 3563 languages and dialects, making the database a vast linguistic resource. The uneven quality of recordings is a drawback, since the database was not created for scientific research. However, the recordings have been used for linguistic research into subjects as rhythm and phonological characteristics [4], developing and testing computer systems to recognize languages [3] and for documenting and reviving rare languages [17].

Languages were chosen for further processing based on recording length and number of native speakers. From the top 300 languages which were analyzed in [22], we selected 239 language to enable us to compare our results with other approaches.

Language recording length makes it possible to extract at least a few thousand segments for each language, up to almost two million for English and Mandarin. To make computation feasible, the number of segments for further processing was restricted to two hundred thousand segments randomly selected from language data.

3 Segmentation

The vast majority of speech processing methods need segmentation of speech signals [5]. Uniform segmentation is used most commonly, but many studies relate to the non-uniform

segmentation of: phones [8, 24], syllables [13], words and other elements [1, 15, 20]. The large variety of segmentation issues determines the multitude of algorithms and the publications which present them. We focused on methods based on wavelet transformation (e.g. [21, 24]).

The continuous nature of speech makes segmentation uncertain. Moreover, various acoustic segments may represent a single phonetic segment and vice versa. In [18] a phone segmentation based on frequency features detected in a speech signal was compared with a segmentation created by human transcribers.

The first stage of our speech analysis is extracting segments corresponding to phones. We used segmentation developed by Ziółko et al. [24]. This spectral method is based on the wavelet packet transformation which splits the speech signal into seven frequency bands. Each fraction is separated by digital low-pass and high-pass filters. Low frequencies have narrow bandwidths and are investigated with a finer resolution, while high frequencies have wide bandwidths, resulting in a lower resolution. The frequency ranges of the seven bands are: 0.5-1 kHz, 1-1.5 kHz, 1.5-2 kHz, 2-3 kHz, 3-4 kHz, 4-6 kHz and 6-8 kHz. In practice the boundaries between these bands overlap because digital filters do not have perfect frequency characteristics. Such speech analysis in the frequency domain corresponds to a perceptual scale.

The role of the segmentation algorithm is to detect significant transitions of energy among the frequency bands. Boundaries of phones are detected based on local changes in energy distribution. This method is universal enough to handle any language. We verified experimentally that having more than seven frequency bands increased the number of segments in comparison with manual segmentation.

Figure 1 is an example of speech segmentation based on energy distribution in seven frequency bands. The upper plot shows the wavelet time-frequency representation of speech signal presented in the lowest part of Fig. 1. The Meyer wavelet of the 11-th order was

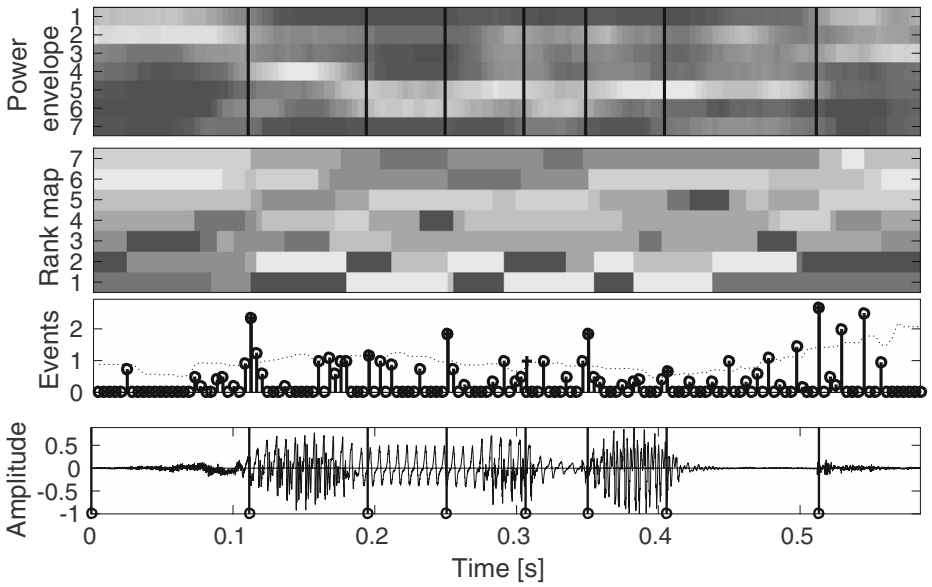

Fig. 1 An example of segmentation based on the time-frequency analysis of speech signal

used. The other two figures show the rank map and event function. The rank map shows the size of energy changes in the frequency bands. The event function presents the global importance of changes in energy distribution.

4 Parametrization

Phones are treated as quasi-stationary segments. We assumed that the majority of phone identity information is concentrated in the centers of the segments. The parameters were calculated for speech segments scaled by the Hamming window to minimize co-articulation effects. Analysis was carried out by applying similar discrete wavelet packets as for segmentation, but with more frequency bands. Phone parameters were calculated as the average energy in eleven frequency bands (see Fig. 2). This way, every extracted phone was characterized by the time stationary vector in the 11th dimensional frequency domain. Details are presented by Ziółko et al. [24]. Such frequency analysis is similar to the commonly used MFCC method. In both approaches, the analysis is carried out on frequency subbands with variable width. The most important difference is the lack of triangular windows and smaller overlapping ranges in our method.

5 Clusterization

The clusterization algorithm involves creating a Gaussian Mixture Model (GMM) to approximate the probability density distribution of phones in an 11-dimensional frequency space. This approach is justified by the common use of GMM in speech modelling. We chose 1024 components (frequently used in other speech applications), which is significantly higher than the expected number of phone representatives in any language. Phone component groups were created by GMM hierarchical clustering. A similar approach to clustering was presented in [7]. Differences between components were calculated as

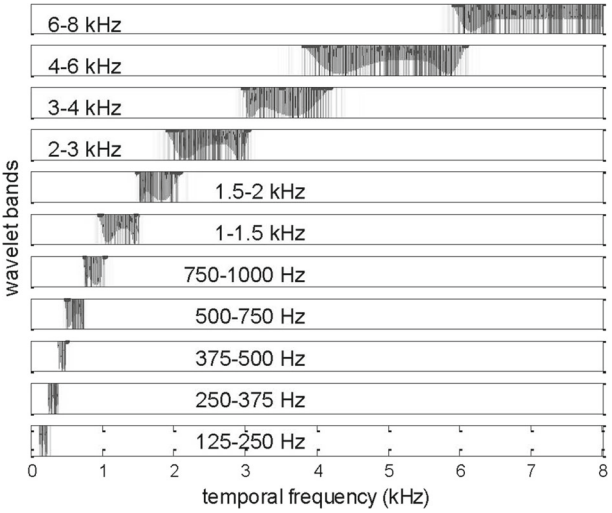

Fig. 2 Frequency bands of Wavelet Packet Decomposition for phones parametrization [24]

Euclidean distances between expected values, Ward's algorithm [23] was then used in a hierarchical clustering procedure.

GMM is associated with the probability density function

$$p(x) = \sum_{k=1}^{K} \alpha_k \mathcal{N}\left(x|\overline{x}_k, \Sigma_k\right), \tag{1}$$

where α_k is the mixture weight and K is the number of components equal to 1024 in our case. The multivariate Gaussian density distribution has the form

$$\mathcal{N}\left(x|\overline{x}_k, \Sigma_k\right) = \frac{1}{\sqrt{(2\pi)^{11}|\Sigma_k|^{1/2}}} \exp\left(-0.5\,(x-\overline{x}_k)^T \Sigma_k^{-1}(x-\overline{x}_k)\right), \tag{2}$$

where the observation $x \in \Re^{11}$ is a cosine transform of a vector representing the energy distribution for a phone and $|\Sigma_k|$ is the corresponding determinant. Cosine transformation allows us to obtain diagonal covariance matrices Σ_k. Finally, the GMM model of phone distribution in the frequency domain is represented by weighting coefficients α_k and the parameters of Gaussian functions: $\overline{x}_k$ and Σ_k^{-1}.

Figure 3 presents the hierarchical clusterization of GMM components for English. The dendrogram shows the dependency of grouping GMM components in clusters and the cut-off point of Ward's distance for 34 phone representatives.

Phone clusterization makes it possible to determine the statistical relationship between phones and phonemes for annotated speech samples. Such experiments showed that pure frequency analysis does not lead to credible mapping between acoustic units (phones) and linguistic transcriptions (phonemes). Left and right context information plays a vital role in accurate phone recognition.

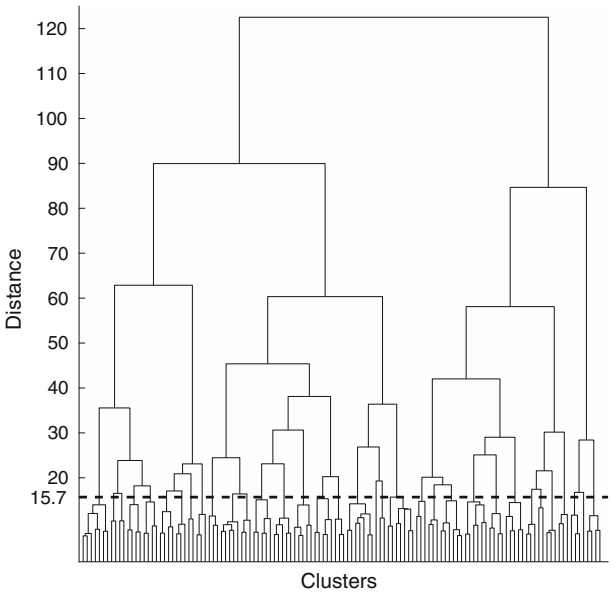

Fig. 3 Results of hierarchical clustering of GMM model for English. The doted line represents the cut-off Ward's distance for 34 phones. For clarity, the bottom part of the dendrogram (with 1024 leaves) is not shown

6 Language differences based on the clustering procedure

The number of clusters c depends on an assumed admissible diversity ρ of elements within the clusters. It decreases if a greater diversity in each cluster is allowed. Figure 4 shows examples of the relationship between the number of clusters and Ward's distance. These plots display the wide range of changes in the number of clusters. A distinctive visual property is the convergence of all the charts for a small and large number of clusters. The most significant differences appear if the number of clusters is in the range typical for the number of phonemes assigned to languages. It is generally assumed that the average number of phonemes for world languages is around 34.

The experimental results characterized by Fig. 4 can be analyzed in many different ways. An important advantage is the ability to precisely approximate experimental data by the equation

$$c(\rho) = a_1 e^{-b_1 \rho} + a_2 e^{-b_2 \rho} \,, \tag{3}$$

where a_1, a_2, b_1, b_2 are parameters chosen separately for each language. We fitted relationship (3) to experimental data for the range of cluster numbers from $c_{min} = 1$ to $c_{max} = 512$. If the number of clusters is equal to the number of Gauss functions (i.a. $c = 1024$), then each cluster contains one element only and the largest distance inside the clusters is equal to 0. This means that all curves shown in Fig. 4 must end at the point: distance = 0 and #Clusters = 1024. There are no differences between languages, so this is not interesting. Model (3) proposed by us is a good representation of experimental data ranging between 1 and 512 clusters. This area is important for language differentiation. Including data for more clusters than 512 would reduce the visibility of differences between languages.

Adjusted R-square statistics was used to verify the mathematical model quality for each language. For the case of mathematical model (3) with four parameters, adjusted R-square statistics for i-th language has the form

$$\overline{R}_i^2 = 1 - \frac{(1 - R_i^2)(J-1)}{J-5} \tag{4}$$

where

$$R_i^2 = 1 - \frac{\sum_{j=1}^{J} (c_{i,j} - c_i(\rho_{i,j}))^2}{\sum_{j=1}^{J} \left(c_{i,j} - \frac{1}{J} \sum_{j=1}^{J} c_{i,j} \right)^2} \tag{5}$$

and $J = 512$ is the number of cluster changes, $c_{i,j}$ is the number of clusters when Ward's distance is not greater than $\rho_{i,j}$, while $c_i(\rho_{i,j})$ is the value of (3) for $\rho_{i,j}$.

The mathematical model (3) for English is characterized by the Root Mean Squared Error RMSE=2.97 and $\overline{R}^2 = 0.9996$. For other languages, the fitting parameters are similar. The worst match was observed for Spanish, we obtained RMSE=6.25 and $\overline{R}^2 = 0.998$.

There are languages which have a low frequency diversity, while in other languages differences between elements in clusters are significantly more noticeable. The relationships between the number of clusters c and the allowed distance ρ for two selected languages are presented in Fig. 5. The examples shown in this figure represent languages having extreme properties in the distribution of phones.

The area

$$A = \int_0^\infty c(\rho)\, d\rho = \frac{a_1}{b_1} + \frac{a_2}{b_2} \,, \tag{6}$$

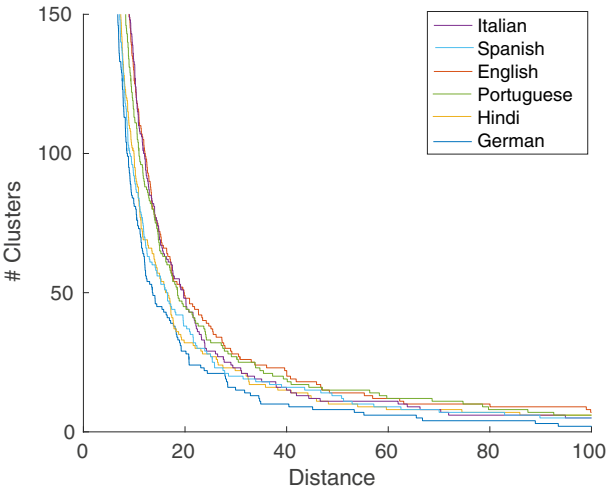

Fig. 4 Number of clusters vs. the cut-off distance (compare with Fig. 3)

under the curve defined by (3) can be taken as the characteristic parameter for each analyzed language. The advantage of this scalar factor is its simple dependence on experimental parameters a_1, a_2, b_1, b_2 characterizing the selected language. Small values of (6) indicate a high decreasing of function (3). In this case, a relatively small change in Ward's distance results in a significant change of cluster number. This means a small variety of articulated phones. Therefore parameter (6) characterizes the frequency diversity of phones. This means that (6) can be used for the ranking of languages.

The clustering procedure starts from 1024 components, because this number of Gauss functions was used to approximate the probability density. The number of clusters decreases as a result of the implementation of Ward's algorithm. A pair of variables is successively obtained: the number of clusters and the maximal Ward's distance between elements within the clusters. This is shown in Fig. 4 for six selected languages.

Assuming that the number of phones is equal to the number of phonemes assigned to the analyzed language, the diversity of phones can be assessed. On the basis of linguistic data it is possible to determine the average value of phonemes for main languages. From data contained in [22] the expected value is slightly above 34 phonemes. For this number of clusters Fig. 5 shows significant variations between languages.

Assuming the number of clusters equal to 34 for all languages, it is possible to systematize them and group languages in terms of similarity. Let set

$$P = \{\rho_i : c_i(\rho_i) = 34\}_{i=1}^{239}, \tag{7}$$

groups characteristic distances for 239 languages being compared. Values of ρ_i depend on a frequency variety of phones. They can be determined directly from the clustering procedure, so they do not depend on the quality of mathematical model (3).

Figure 6 presents the flowchart of calculations provided for each language separately. Most of the calculations: DWT parametrization, GMM training, clustering and curve fitting is done using built-in MATLAB functions. Implementation of speech segmentation algorithm was obtained from authors of [24].

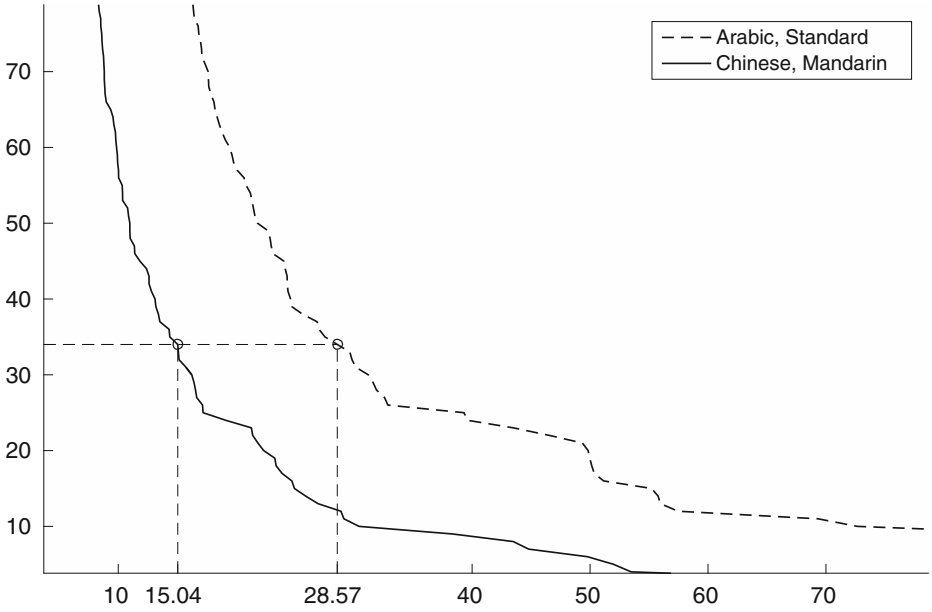

Fig. 5 Number of clusters as a function of the cut-off Ward's distance for Arabic and Mandarin

7 Experiments

Both indicators (6) and (7) can be used to assess the diversity of phones for the analyzed languages. The indicator (7) is calculated directly from the results of the clusterization. However, it is sensitive to local deviations and therefore the indicator (6) seems to be more accurate for the ranking of languages.

Table 1 presents two indicators which characterize the chosen 50 languages. The first indicator is defined by (6) and its values are presented in second column. The second indicator is Ward's distance assuming that the number of clusters is equal to 34. Values of this indicator are defined by (7) and are shown in the third column of Table 1. Both indicators are measures of phone diversity in the frequency domain, therefore they should be correlated. For the 239 analyzed languages the correlation coefficient is equal to 0.72. The languages are ordered from the highest to the lowest value of indicator (6). It means that languages which have relatively major differences in articulation are shown at the top of Table 1. This group includes Arabic and Punjabi. In contrast Mandarin is characterized by the lowest variation in phones articulation.

The other four columns present coefficients of mathematical model (3). This model is the sum of two exponential functions. The initial values of the first functions are approximately ten times higher (a_1 in relation to a_2), but their decay rates are approximately six times higher (b_1 in relation to b_2). As a result, the second components of the model (3), determined by parameters a_2 and b_2, have a greater impacts on modeling effects for $\rho > 10$.

The last two columns of Table 1 present RMSE and adjusted R-square statistics (4). The value of index (4) is equal to 1 if the mathematical model provides a perfect approximation. The data presented in the last column of Table 1 indicates very good usability of model (3). The next to last column presents RMSE values. For all languages these errors concern the

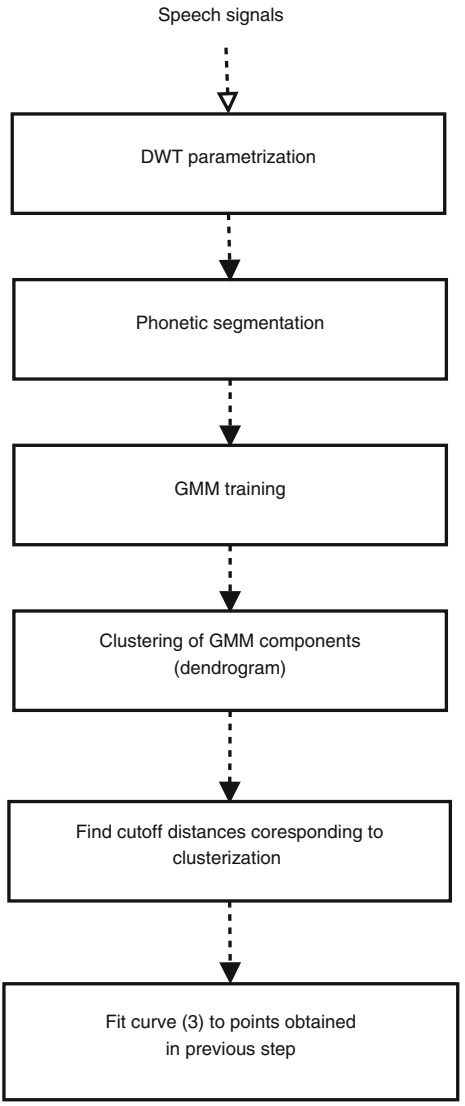

Fig. 6 Flowchart of calculations to determine the mathematical model (3) for tested language

number of clusters c which vary from 1 to 512. This index is more sensitive and it makes it possible to differentiate modeling efficiency when all results are very good.

If we assume that the number of clusters is the same as the number of phonemes, then we can suppose that each cluster corresponds to a certain phoneme. To verify this hypothesis, experiments have been provided for languages taken from corpora with hand annotations. It appears that only 20% of phonemes were correctly allocated to clusters [12]. This observation is not surprising and was firstly noted around 60 years ago (e.g. [16]). Now, the hidden Markov models are used in automatic recognition systems to overcome these difficulties.

Table 1 Ranking of languages according to phones diversity in the frequency domain

Language	Equation (6)	Equation (7)	a_1	b_1	a_2	b_2	RMSE	$\overline{R}^2$
Pular	8241	22.6	3586	0.7	338	0.11	3.62	0.999
Croatian	8093	21.91	3212	0.62	294	0.1	4.23	0.999
Arabic	8029	28.57	1757	0.37	199	0.06	3.25	1
Punjabi	7703	28.25	2070	0.45	229	0.07	5.25	0.999
Italian	7697	23.2	2079	0.42	183	0.07	4.27	0.999
Vietnamese	7440	19.8	3827	0.79	320	0.12	3.97	0.999
Marathi	7431	20.75	3108	0.64	272	0.1	4.84	0.999
Javanese	7335	26.09	2076	0.47	227	0.08	3.56	0.999
Swedish	7153	23.62	2658	0.61	285	0.1	4.1	0.999
Urdu	7109	26.47	1613	0.37	154	0.06	4.34	0.999
Spanish	7083	21.56	3283	0.73	300	0.12	6.25	0.998
Hungarian	7052	22.68	2596	0.6	291	0.11	4.57	0.999
Romanian	7014	19.28	2977	0.68	309	0.12	4	0.999
Japanese	6986	27.05	1746	0.44	223	0.07	3.52	0.999
Ukrainian	6963	22.66	2669	0.66	318	0.11	4.37	0.999
English	6957	26.12	1693	0.42	195	0.07	2.97	1
Polish	6946	27.43	1949	0.46	193	0.07	3.7	0.999
Serbian	6863	25.36	2154	0.58	351	0.11	5.32	0.999
Bengali	6812	17.56	3264	0.73	302	0.13	4.24	0.999
Armenian	6809	21.37	2140	0.54	311	0.11	4.79	0.999
Somali	6806	21.85	2581	0.63	284	0.11	4.11	0.999
Portuguese	6675	24.12	1869	0.49	228	0.08	3.7	0.999
Turkish	6559	20.3	2833	0.69	268	0.11	4.24	0.999
French	6449	23.74	2008	0.52	224	0.09	4.69	0.999
Dari	6444	17.56	3367	0.82	316	0.14	4.56	0.999
Korean	6414	21.66	2352	0.61	251	0.1	5.15	0.999
Greek	6341	21.71	2136	0.56	246	0.1	5.34	0.999
Georgian	6314	19.5	2405	0.63	268	0.11	3.5	0.999
Slovak	6280	23.62	2033	0.57	246	0.09	4.55	0.999
Dutch	6249	18.76	2249	0.63	342	0.13	4.11	0.999
Turkmen	6196	19.55	2651	0.7	276	0.11	3.95	0.999
Lithuanian	6179	18.29	2297	0.57	201	0.09	3.6	0.999
Indonesian	6175	20.78	2539	0.69	300	0.12	5.66	0.999
Hindi	6130	19.18	2082	0.56	240	0.1	4.38	0.999
Pashto	6109	18.73	2686	0.72	296	0.12	4.34	0.999
Hebrew	6080	18.36	2540	0.67	267	0.12	3.83	0.999
Bulgarian	5975	22.21	2115	0.59	203	0.09	3.71	0.999
Russian	5935	17.56	2307	0.62	254	0.11	3.58	0.999
Kyrgyz	5925	18.88	2485	0.67	250	0.11	4.05	0.999
Farsi	5903	18.04	2560	0.71	280	0.12	3.56	0.999
Kannada	5887	18.65	2312	0.63	245	0.11	3.74	0.999
German	5823	18.4	2328	0.65	261	0.12	4.05	0.999
Danish	5753	18.39	1881	0.5	157	0.08	2.36	1

Table 1 (continued)

Language	Equation (6)	Equation (7)	a_1	b_1	a_2	b_2	RMSE	$\overline{R}^2$
Tamil	5640	17.63	2105	0.61	236	0.11	3.06	1
Finnish	5575	18.75	2160	0.64	228	0.1	4.09	0.999
Telugu	5504	14.52	2989	0.84	313	0.16	4.8	0.999
Czech	5398	14.57	2774	0.8	305	0.16	4.07	0.999
Mandarin	5048	15.04	2550	0.79	271	0.15	3.7	0.999

8 Conclusions

Our research was inspired by work carried out by Atkinson [2] in which he compared the phoneme diversity for 504 languages. Our main motivation was to find an acoustic similarity measure between languages that can lead to language taxonomies. We supposed that this measure could be used to verify Atkinson's hypothesis about the presence of a founder effect in world languages. The comparison of the language ranking obtained by us with the results of Atkinson's work does not confirm his hypothesis. Our experiments support the views of Atkinson's adversaries, claiming that various factors conditioned by historical processes have a decisive influence on the diversity of articulation. These phenomena have a major impact on the evolution of languages. Their relatively high rate of change is clearly signalled in other studies, e.g. [9].

The data we obtained can be correlated with the geographical location of languages; additionally, there may be other phenomena which have a significant impact on the size of the differences in phone pronunciation. This direction of research could lead to interesting conclusions.

Our main conclusions arise from the analysis of data showing the relationship between the number of clusters and their internal differentiation. There are no clear isolated clusters in the frequency space. This makes it possible to fine-tune the continuous curve (3). However, Fig. 5 shows the existence of 24 visible isolated clusters for Chinese and fewer isolated clusters for Arabic.

The method of clustering we used is frequently applied in various types of scientific research. A great simplification is the availability of ready-made computer programs. Calculating the differences between components remains an open question, better methods than the Euclidean distance may exist.

Frequency analysis of phones is not sufficient to reliably determine phonemes in speech recognition systems. Although both types of frequency analyzers, the sense of hearing and electronic devices - operate efficiently, they cannot remove distortions from speech. The human brain and computer analysis (supported by trained models) play a highly important role in speech recognition.

We used the frequency variety of phones to rank the order of languages. We ranked languages from those where frequency diversities between phones are significant to languages where these differences are significantly smaller.

Major differences in the articulation of phones may involve languages spoken by non-native speakers i.e. people with diverse cultural backgrounds. Secondly, it seems that major differences in articulation make learning foreign languages easier.

Smaller differences between phone articulation may be due to fast speech. Schupert et al. [19] verified the hypothesis that differences in speech tempo are the main reason why spoken Danish is so difficult to understand for Norwegians and Swedes. Differences between Danish and Swedish, shown in Table 1, support this conclusion.

Data for all 239 tested languages is available from http://www.dsp.agh.edu.pl/_media/pl:research:language_ranking.pdf.

Acknowledgements The project was funded by the National Science Centre granted on the basis of the decision DEC-2011/03/B/ST7/00442.

References

1. Amirgaliyev Y, Hahn M, Mussabayev T (2017) The speech signal segmentation algorithm using pitch synchronous analysis. Open Computer Science 7(1):1–8
2. Atkinson QD (2011) Phonemic diversity supports a serial founder effect model of language expansion from Africa. Science 332(6027):346–349
3. Castaldo F, Dalmasso E, Laface P, Colibro D, Vair C (2008) Politecnico di torino system for the 2007 nist language recognition evaluation. ISCA
4. Easterday S, Timm J, Maddieson I (2011) The effects of phonological structure on the acoustic correlates of rhythm. ICPhS XVII pp 623–626
5. Fukada T, Bacchiani M, Paliwal KK, Sagisaka Y (1996) Speech recognition based on acoustically derived segment units. In: Proceedings of the 4th international conference on spoken language, 1996. ICSLP 96, vol 2, pp 1077–1080. IEEE
6. Global recordings network. http://globalrecordings.net
7. Goldberger J, Roweis ST (2004) Hierarchical clustering of a mixture model. In: Advances in neural information processing systems, pp 505–512
8. Hoang DT, Wang HC (2015) Blind phone segmentation based on spectral change detection using Legendre polynomial approximation. J Acoust Soc Am 137(2):797–805
9. Holman EW (1996) Quantitative properties of the evolution and classification of languages. J Classif 13(1):27–56
10. Jain AK (2010) Data clustering: 50 years beyond k-means. Pattern Recogn Lett 31(8):651–666
11. Jansen A, Church K (2011) Towards unsupervised training of speaker independent acoustic models. In: Proceedings of the interspeech, pp 1693–1692
12. Kacprzak S, Mąsior M, Ziółko M (2016) Automatic extraction and clustering of phones. In: 2016 Signal processing: algorithms, architectures, arrangements, and applications (SPA), pp 310–314
13. Laleye FA, Ezin EC, Motamed C (2017) Automatic text-independent syllable segmentation using singularity exponents and rényi entropy. Journal of Signal Processing Systems 88(3):439–451
14. Ma B, Li H, Lee CH (2005) An acoustic segment modeling approach to automatic language identification. In: 9th european conference on speech communication and technology
15. Martínez-González B, Pardo JM, Echeverry-Correa JD, San-Segundo R (2017) Spatial features selection for unsupervised speaker segmentation and clustering. Expert Syst Appl 73:27–42
16. Peterson GE, Barney HL (1952) Control methods used in a study of the vowels. J Acoust Soc Am 24(2):175–184
17. Rybka K (2015) State-of-the-art in the development of the Lokono language
18. Scharenborg O, Wan V, Ernestus M (2010) Unsupervised speech segmentation: an analysis of the hypothesized phone boundaries. J Acoust Soc Am 127(2):1084–1095

19. Schüppert A, Hilton NH, Gooskens C (2016) Why is Danish so difficult to understand for fellow scandinavians? Speech Comm 79:47–60
20. Singh R, Raj B, Stern RM (2002) Automatic generation of subword units for speech recognition systems. IEEE Transactions on Speech and Audio Processing 10(2):89–99
21. Tan BT, Lang R, Schroder H, Spray A, Dermody P (1994) Applying wavelet analysis to speech segmentation and classification. In: Wavelet applications, vol 2242, pp 750–762. International society for optics and photonics
22. Wang CC, Ding QL, Tao H, Li H (2012) Comment on phonemic diversity supports a serial founder effect model of language expansion from Africa. Science 335(6069):657
23. Ward JH Jr (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301):236–244
24. Ziółko M, Gałka J, Ziółko B, Drwięga T (2010) Perceptual wavelet decomposition for speech segmentation. In: Proceedings of the interspeech, pp 2234–2237

Mariusz Ziółko received his M.Sc. in electrical engineering in 1970, his Ph.D. in automatic control in 1973, and his D.Hab. in 1990, all from the AGH University of Science and Technology, Kraków, Poland. He is currently working as Professor at the AGH University of Science and Technology. He has authored or coauthored more than 150 scientific papers published in journals including IEEE Transactions on Automatic Control, Mathematical Biosciences, Theoretical Population Biology, Functional Ecology, Applied Numerical Mathematics, Kidney International and IEEE Transactions on Fuzzy Systems. His research interests include speech technology, signal processing, modeling of biomedical processes, and applications of mathematics.

Stanisław Kacprzak received his M.Sc. in computer science from the Technical University of Lodz in 2011. He is currently working on his PhD at the AGH University of Science and Technology in Kraków. His research interests include speech technology, mainly speaker and language recognition.