

Diversity of phones pronunciation between world languages

Mariusz Ziółko, Stanisław Kacprzak

Plzen, 11 – 12 October 2016

The project was funded by the  NATIONAL SCIENCE CENTRE
POLAND
on the basis of the decision number DEC-011/03/B/ST7/00442

INTERSPEECH'15

The voice of authority: > 10 000 h of speech data
It means 3 years and 30 mln \$

Is there a Problem ?

- Modern speech processing requires >10,000h of speech data
 - e.g. speech recognition, speaker recognition, speech synthesis, speech coding
- Large databases of spoken language have been collected for >20 years
 - by commercial organisations and by publicly funded research projects
 - speaker names and other identifying metadata are always (?) suppressed
- State-of-the-art speaker recognition achieves error rates (EER) <1%
- Recent research has found correlates of medical, psychological and behavioural conditions in speech
 - e.g. depression, drug and alcohol use
- With current and future system capabilities, ...
 - ... **Should we be concerned?**

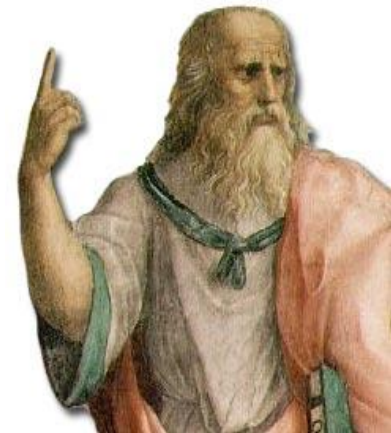


Phone versus phoneme (Crystal 1971, p. 180)

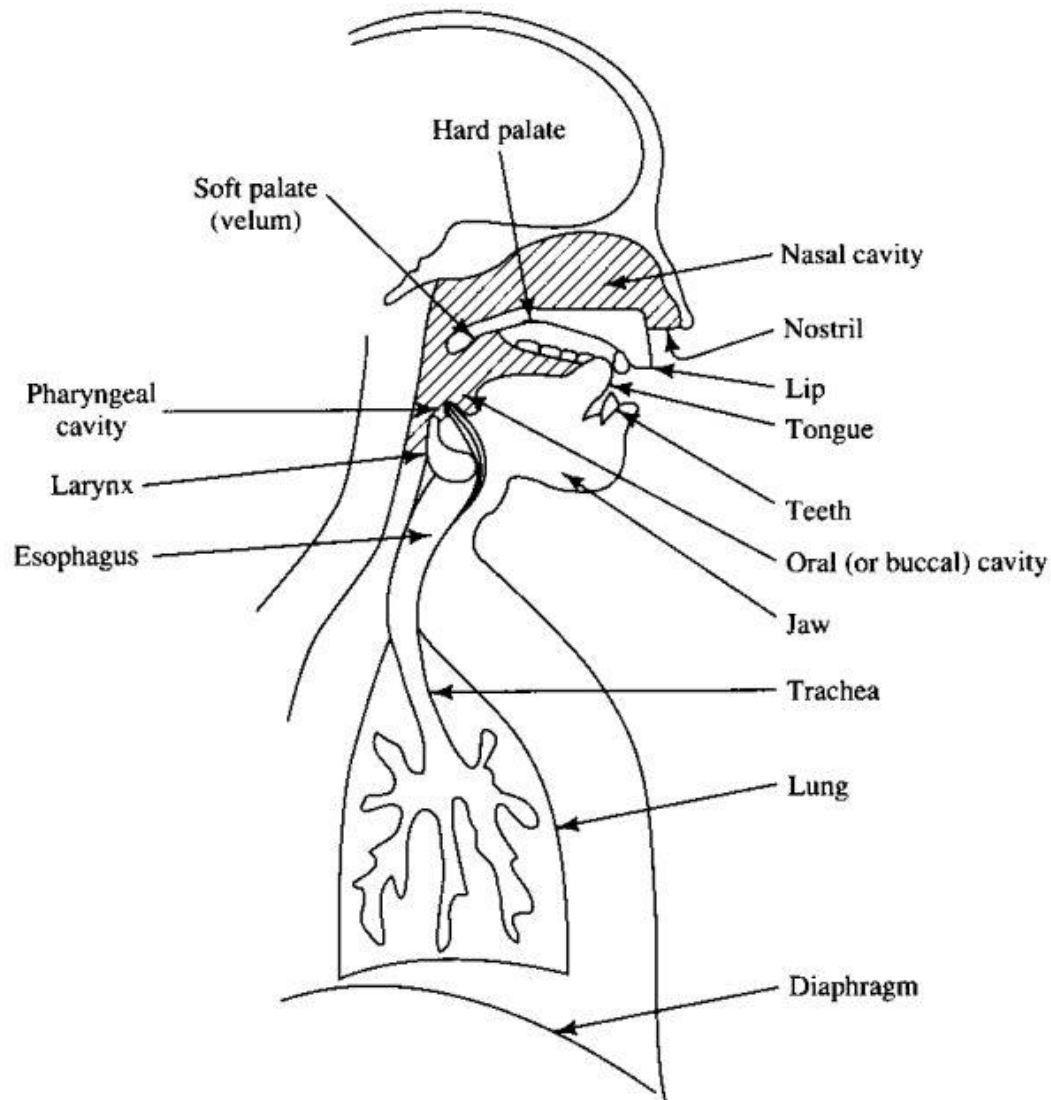
„In phonetics and linguistics, the word **phone** may refer to any speech sound considered as a physical event without regard to its place in the phonology of a language”.

„In contrast, a **phoneme** is a set of phones or a set of sound features that are thought of as the same element within the phonology of a particular language”.

Small number of languages have transcribed training data corpora. Our motivation to develop the universal method of automatic extraction of phones from non-annotated speech is a need to compare the phones of huge number of languages (more than 200).




Vocal tract



From physical point of view, the **speech** signal **is strongly distorted** by the individual's characteristics such as: sex, age, intonation, and emotional state. Additionally distortions in the form of co-articulation brings inertia of voice track, a significant influence of neighboring phones.

How in spite of **many distortions**, the speech is **accurately analyzed** by the **human sense** of hearing, and how speech signal is efficiently process by **technical devices**?

Data sources



The screenshot shows the homepage of the Global Recordings Network (GRN). The header includes the GRN logo, the tagline "Telling the story of Jesus in every language", and a search bar. Navigation menus for Home, Listen or Download, Resources, News, Be Involved, About Us, and Contact Us are visible. A main section titled "Spoken Languages of the World" contains text about 12,000 spoken languages and a search interface with radio buttons for "Begins with", "Ends with", and "Contains". A "Find out more" section with social media icons is also present. The footer lists various languages like Thai, Bahasa Indonesia, Deutsch, etc.

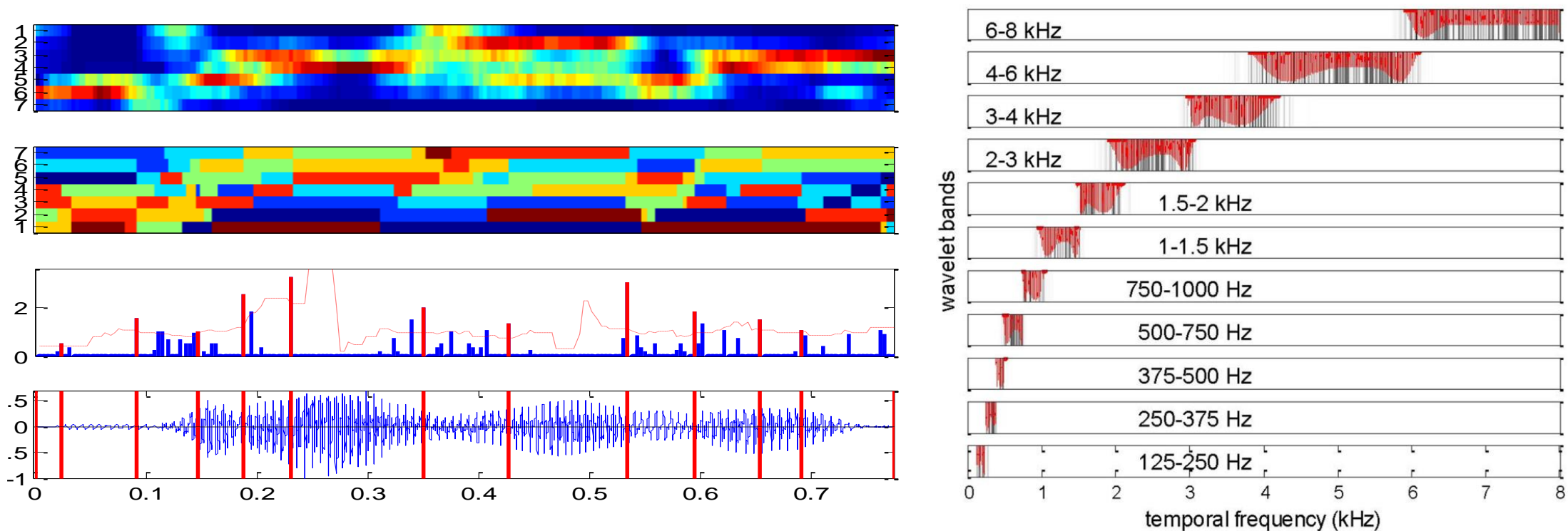
<http://globalrecordings.net>



The screenshot shows the homepage of the Speech Samples website. The header features the title "Speech Samples" and a navigation menu with Home, About Research, News, Partners, Feedback, and SpeechSamples DB. The main content area is titled "Speech Samples Research Project" and describes the goal of investigating the origin and expansion of modern languages. It lists three key objectives: collection of speech recordings, investigation of acoustical differences, and creation of a natural languages taxonomy. A prominent "SPEECH SAMPLE" button is visible. A "Latest news" sidebar on the right contains two items: "Website is operational!" and "SpeechSamples webpage". The footer includes social media links for Twitter, Facebook, and YouTube, and mentions the site is powered by AGH Signal Processing Group.

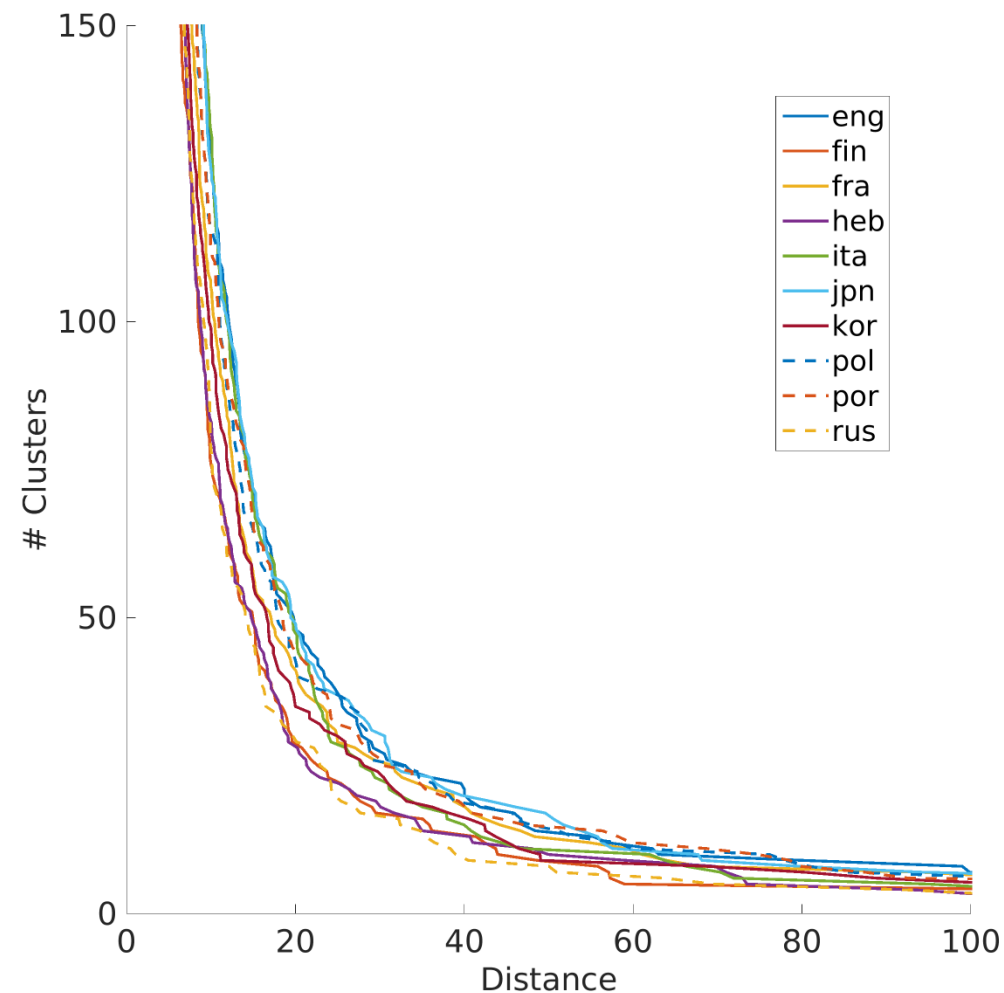
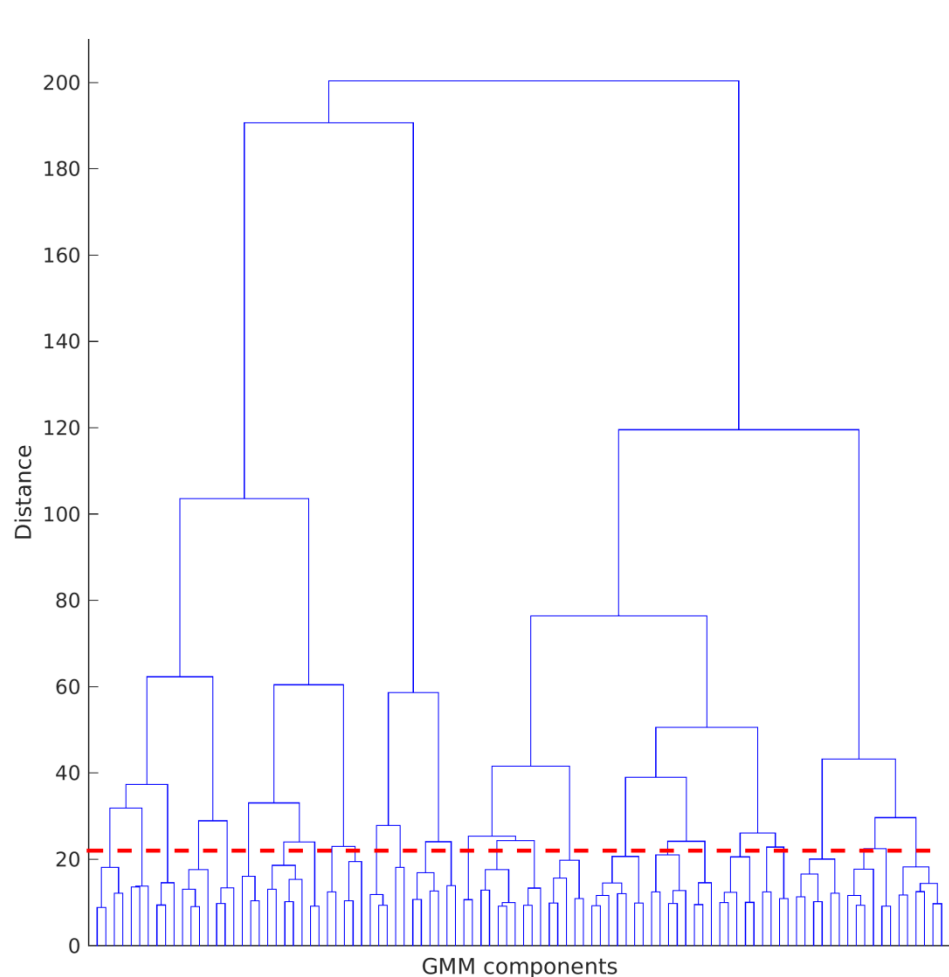
<http://speechsamples.agh.edu.pl>

Automatic phone segmentation



It was experimentally verified, that the greater number than 7 frequency bands increases the number of segments in comparison with the manual segmentation. The phones boundaries were fixed in places of relatively large changes in the energy distribution between the frequency bands. Average duration of obtained segments was 73 ms. The phone parameters were calculated as an average energy in 11 frequency bands.

Number of clusters (sets of phones) in dependency of acoustic differences



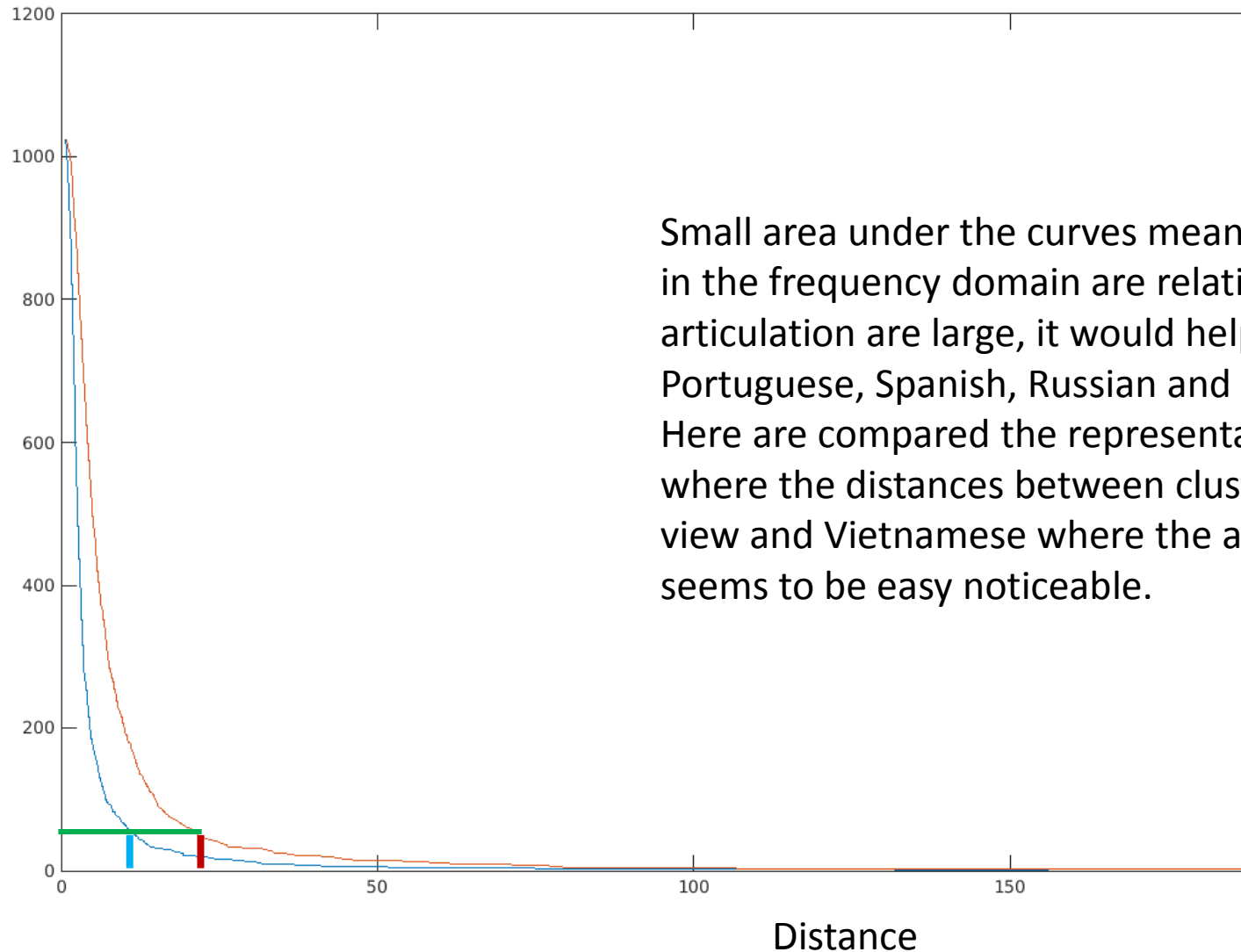
Analysis of the frequency properties results in 20% of correct phoneme recognitions only.

Comparision of languages

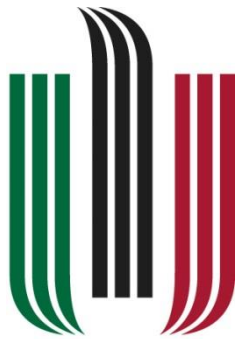
1	Bulgarian	2703		24	Shona	3966		47	Tamil	4774
2	Hungarian	3140		25	Kyrgyz	3994		48	German	4777
3	Gujarati	3219		26	Slovak	3996		49	Gikuyu	4801
4	Dholuo	3231		27	Sindhi	4053		50	Georgian	4814
5	Turkish	3337		28	Assamese	4053		51	English	4827
6	Croatian	3425		29	Konkani	4083		52	Hiligaynon	4872
7	Korean	3461		30	Kurdish	4131		53	Tagalog	4910
8	Uzbek	3470		31	Serbian	4136		54	Armenian	4934
9	Slovene	3488		32	Lao	4163		55	Russian	4972
10	Bengali	3498		33	Bemba	4186		56	Mongolian	4989
11	Lingala	3537		34	Czech	4212		57	Dutch	5014
12	Tibetan	3568		35	Pashto	4245		58	Kituba	5030
13	Finnish	3591		36	Polish	4253		59	Gilaki	5030
14	Urdu	3654		37	Turkmen	4259		60	Spanish	5034
15	Sotho	3679		38	Italian	4266		61	Portuguese	5067
16	Cebuano	3687		39	Igbo	4468		62	Lithuanian	5242
17	Shan	3734		40	Japanese	4544		63	Amharic	5287
18	Romanian	3743		41	Kazakh	4574		64	Kabyle	5309
19	Swedish	3777		42	Ukrainian	4592		65	Thai	5445
20	Uyghur	3878		43	Sinhala	4605		66	Telugu	5445
21	Malayalam	3894		44	French	4624		67	Fon	5516
22	Khmer	3919		45	Marathi	4701		68	Danish	6609
23	Balochi	3921		46	Kiche	4714		69	Vietnamese	7684

Comparision of Hungarian and Vietnamese languages

Clusters



Small area under the curves means that differences between clusters in the frequency domain are relatively small. If the differences in articulation are large, it would help foreigners to learn such languages. Portuguese, Spanish, Russian and English belong to these group. Here are compared the representatives of both groups. Hungarian, where the distances between clusters are small from acoustic point of view and Vietnamese where the acoustic differences between clusters seems to be easy noticeable.



AGH



Thank you

