Krajowa Konferencja Zastosowań Matematyki w Biologii i Medycynie

Sandomierz, 5th–9th September 2016

MULTILANGUAGE MODELING OF PHONES

Stanisław Kacprzak, Mariusz Ziółko

AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Kraków,
{skacprza, ziolko}@agh.edu.pl

ABSTRACT

The wavelet analysis was used for speech segmentation and parameterization. The obtained frequency parameters of phones were grouped using the Gaussian Mixture Model and the hierarchical clustering. The relationship between the number of clusters and a maximum distance between their centers was approximated by the sum of two exponential functions. The research was conducted for 245 world's languages. In this way, for each language four parameters were received. A comparison of these parameters allows to search for acoustic similarities between phones of different languages.

INTRODUCTION

Speech is the main way of communicating between people. Our speech ability to communicate provides evidence of abstract thinking, constitutes a proof of humanity and distinguishes people from other species. Scholars estimate that speech appeared 30 000 year ago and the number of living languages varies between 6000 and 7000. Speech is a symbol of humanity and has huge importance for the development of civilization. Today, however, it is possible to talk with a computer. Sometimes computers make it so excellent that it is hard to realize that our interlocutor is not a human being.

The speech signal is generated by a voice track. Conversion of speech acquired from the microphone into the corresponding letter transcription is a key issue for speech technology and is called speech recognition. 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 significant distortions in the form of co-articulation brings inertia of voice track. It means that, there is a significant influence of neighboring phones on phone articulation. All these phenomena strongly impact on the physical properties of speech signal. Therefore, 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, are reasonable questions.

The process of speech understanding starts with the time-frequency analysis. The human ear is a frequency inverter that converts an acoustic signal in the form of air vibrations into the frequency dependent electrical impulses delivered to the brain through the nervous system. A plurality of impacts deforming the generation of speech make the result of ear work not precise in relation to the final effect of the correctly received speech signal content. The brain that interprets signals obtained from ears is able to pick out the relevant content from distorted information. Because of the fundamental importance of the brain, we have to treat it as an integral and the most important

part of the hearing organ. Roughly we guess how processes run in the brain and at the same time we admire the brain efficiency.

Computer algorithms for speech recognition were suggested decades ago and are improved systematically. The major components of the computer analysis are not changed and in the general overview the speech recognition systems use two characteristic procedures of computer analysis. They sometimes differ in details but fulfill the same function. The first procedure is based on frequency analysis using the mel-frequency cepstral coefficients (MFCC). The mel scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another. A classical formula to convert f [Hz] into mel scale is

$$f_m = 2595 \log_{10}(1 + f/700). \tag{1}$$

Therefore both, the technical devices and the human ears, use the frequency analysis as the first stage of the speech signal processing. The frequency representations of speech signal are created by computers usually for 30 [ms] long segments. So short fragments of speech are hardly recognizable by people and seem to be sounds like crackles. To increase the probability of accurate assignment of such a short segment of a speech to the uttered phone, computers compare the results of frequency analysis of the adjacent segments of speech. This is the second important stage of speech recognition systems realized by the Hidden Markov Model (HMM). It appeared that the probability of accurate recognition not a single, but a sequence of consecutive segments is already big enough to obtain an efficient speech recognition systems. A similar phenomenon can be observed in human perception of short (more than 100 [ms]) speech signals.

The acoustic analysis of multi-linguistic speech will provide answers to the following question: which phonemes are used in different parts of the world and what are their individual features. The comparison between languages results in interesting conclusions. Our analysis has covered 70 languages spoken in Europe. The majority of the European population use more than 20 consonants but only a few vowels. The largest group of people use 5 vowels and 36 consonants but there is one language with 84 consonants. Most of European languages have from 5 to 6 vowels and 24 to 25 consonants.

SPEECH SEGMENTATION AND PARAMETRIZATION

The extracting of acoustic segments corresponding to phones was the first stage of our experiments. The algorithm used for segmentation was developed by Gałka [1]. The spectral method based on the Wavelet Packet Decompoition (WPD) was used to split speech into 11 frequency bands (see Fig.1). The representation of a single phone $\{s(n)\}_n$ was chosen as a vector of average energy in 11 frequency bands. Each fraction of

$$WPD = \left\{ \{d_{m,n}\}_n, \{d_{m-1,n}\}_n, \ldots, \{d_{m-10,n}\}_n \right\} \tag{2}$$

was separated by digital filters, low-pass μ_n and high-pass η_n, applying the iterative procedure

$$c_{m,n} = \sum_l \mu_{n-l} c_{m+1,l} \tag{3}$$

$$d_{m,n} = \sum_l \eta_{n-l} c_{m+1,l} \tag{4}$$

which started from resolution level $m + 1$ by substitution

$$c_{m+1,n} \leftarrow s(n). \tag{5}$$

The low frequencies have narrow bandwidths and are investigated with a finer resolution, while the high frequencies have wide bandwidths, what results in a lower resolution. Such speech analysis in frequency domain corresponds to a perceptual scale (1).

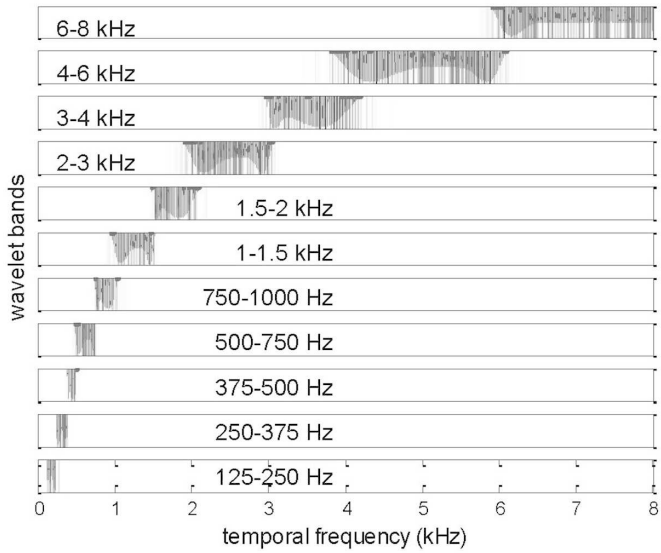

Figure 1. Frequency bands of Wavelet Packet Decomposition for phones analysis [1].

The collected recordings are segmented into elementary units. The role of the segmentation algorithm is to detect significant transitions of the energy among the frequency subbands. Boundaries of phones were detected based on changes in energy distribution between the frequency bands [1]. This methodology provides accurate segmentation and is based on exploration of local changes in energy distribution in time-frequency speech spectrum. Such reduction of data was an essential issue for the further analysis to avoid the computational complexity. Fig.2 is an example of parametric representation of speech segments as energy in mentioned frequency subbands. The method is universal enough to handle any language.

We assumed that the most of phone identity information is concentrated in the center of segment. To minimize the co-articulation effects, the final parameters $x \in R^{11}$ were calculated for speech segments $\{s(n)\}_{n=1}^{N}$ scaled by the Hamming window

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \tag{6}$$

where N is the width of window (number of samples). Phones were described as quasi-stationary processes consisting of stages

$$x = \left[\sum_n d_{m,n}^2 \ \cdots \ \sum_n d_{m-10,n}^2\right] \in R^{11}. \tag{7}$$

PHONES MODELING

Speech parameterization involves the representation of its spectrum in a way that can effectively represent the most relevant information. The determination of the acoustic groups (identified with phones) was carried out by cluster analysis for speech segments. Algorithm consisted on creating the Gaussian Mixture Model (GMM) for phones density in the frequency domain

$$p(x) = \sum_{k=1}^{K} \alpha_k \mathcal{N}\left(x|\overline{x}_k, \Sigma_k\right), \tag{8}$$

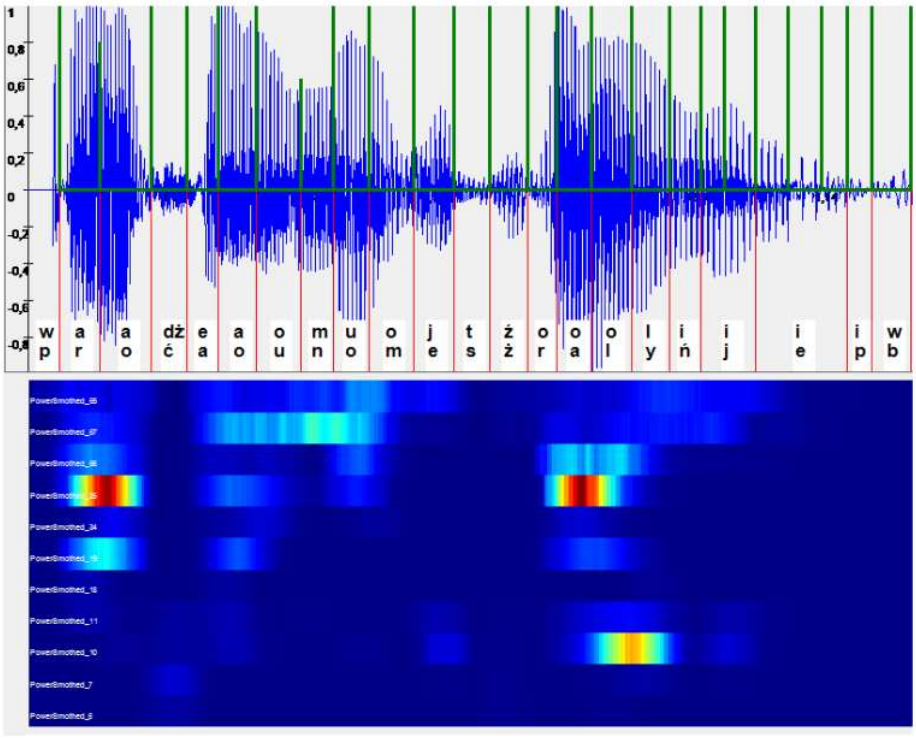

Figure 2. Speech segmentation and parameterization based on the wavelet time-frequency analysis.

$$\sum_{k=1}^{K} \alpha_k = 1, \text{ and } 0 \leq \alpha_k \leq 1, \tag{9}$$

$$\mathcal{N}(x|\overline{x}_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_k|}} \exp\left(-\frac{1}{2}(x-\overline{x}_k)^T \Sigma_k^{-1}(x-\overline{x}_k)\right), \tag{10}$$

where $\overline{x}_k$ is mean vector of k Gaussian component, Σ_k is covariance matrix and K is number of components. Vector parameters $\overline{x}_k$, Σ_k and weights α_k were estimated by Expectation-Maximization (EM) algorithm [2] for the set of phones represented by parameters (7). We chose $K = 1024$ (commonly used in other speech application), which is much larger than expected number of phones in any language. GMM phones component groups were achieved by hierarchical clustering. Similar approach to clustering GMM components has been presented by Goldberg [3]. Differences between components were calculated as the Euclidean distances between expected values of components with the Wards's method (minimum variance algorithm) [4].

Fig.3 presents results of hierarchical clusterization of GMM components. This dendrogram was used to obtain clusteriazations in dependency of the cut-off point. Increasing the allowable distance ρ between the clusters (dashed line in Fig.3 moves to the top) the number of clusters r is reduced. These dependencies for the selected 10 of the world's languages are shown in Fig.4. Each of the experimentally obtained curves can be approximated by function

$$r(\rho) = a_1 \exp(b_1 \rho) + a_2 \exp(b_2 \rho). \tag{11}$$

The mean matching coefficient reached value 0.998 for 245 languages. This means that approximation (11) can be choose with high precision for each language to represent its phones diversity by four parameters: a_1, b_1, a_2, b_2. Such representation (examples for 10 languages are presented in Table 1) allows for further comparison of languages.

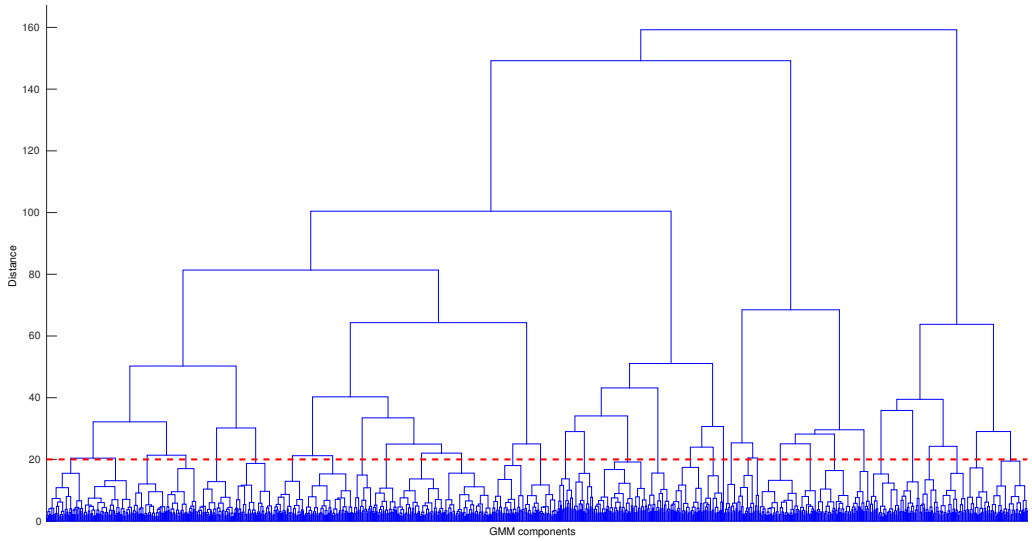

Figure 3. Results of hierarchical clustering (example for Polish language).

In [5] we suggest to choose the characteristic number of phones for each language, based on knee point for the curves presented in Fig.4. More or less in this area, which is slightly below 30 clusters, are the largest deviations between curves presented in Fig.4. Clusters group together these speech segments, which in acoustic terms have relatively small differences and can be identified with phones. Under such assumption, Fig.4 shows that separation of 30 or less phones gives significant acoustic differences between them. So it is easier to distinguish phones from each other. Assuming over 30 phones, differences in their articulation become difficult to detect. This dependence occurs for each language and can be considered as a result of the human perception. This observation is in accordance with the rules of the transcription for most of the world's languages.

Table 1. *Language parameters for 10 of 245 languages* .

Language	a_1	b_1	a_2	b_2
English	1860	-0.42	172	-0.06
Finnish	1669	-0.44	53	-0.02
French	1798	-0.42	106	-0.04
Hebrew	2068	-0.52	122	-0.06
Italian	2454	-0.46	214	-0.08
Japanese	1812	-0.41	166	-0.06
Korean	2084	-0.51	149	-0.06
Polish	1870	-0.40	107	-0.04
Portuguese	1744	-0.41	129	-0.05
Russian	2227	-0.54	132	-0.07

CONCLUSIONS

Before the computer era, linguists have created the taxonomy for the world's languages. They adopted criteria that do not require the tedious calculations. Currently, the technical possibilities allow us to compare languages, search for their similarities and group them on the basis of an enormous amount of data in the form of written text or recorded speech. In this paper we proposed

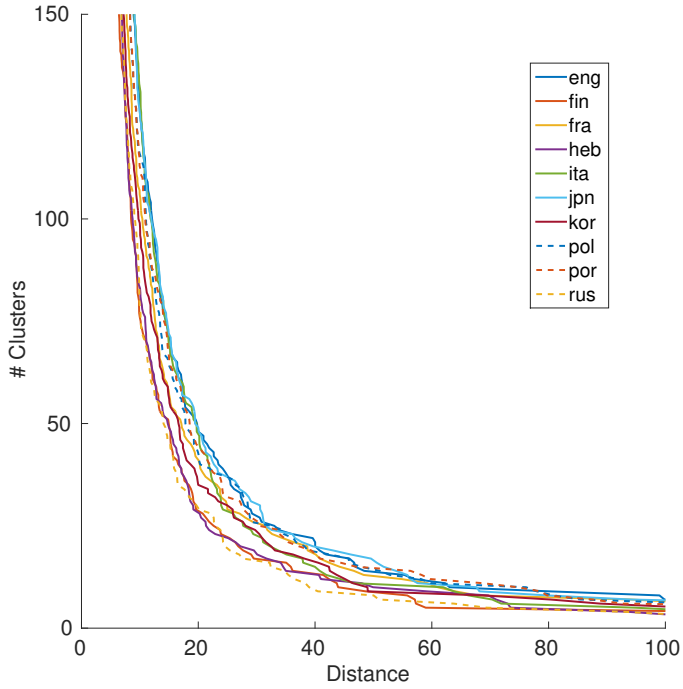

Figure 4. Number of clusters r in dependency of the cut-off point ρ (example for 10 languages).

analysis of long hours of recordings for several hundred of the world's languages. The essence of the proposed method of analysis in the frequency domain is to search acoustic variations of phones in spoken languages. The proposed mathematical model allows to characterize the diversity of phones articulation using four parameters only. This tool will be used to create a language taxonomy based on a comparison of the speech articulations.

ACKNOWLEDGMENTS

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

REFERENCES

[1] J. Galka and M. Ziółko: *MELECON 2008-The 14th IEEE Mediterranean Electrotechnical Conference*, 2008.

[2] A.P. Dempster, N.M. Laird, and D.B. Rubin: *Maximum likelihood from incomplete data via the EM algorithm*, J. R. Stat. Soc. Series B Methodol. **39(1)** (1977), 1–38.

[3] J. Goldberger and S.T. Roweis: *Advances in neural information processing systems 17* (L. K. Saul and Y. Weiss and L. Bottou, ed.), MIT Press, 2004, pp. 505–512.

[4] Jr. Ward and H. Joe: *Hierarchical grouping to optimize an objective function*, J. Amer. Statist. Assoc. **58(301)** (1963), 236–244.

[5] S. Kacprzak, M. Mąsior, M. Ziółko, and et al: *Automatyczna ekstrakcja i klasteryzacja głosek w sygnale mowy dla wielojęzykowej analizy porównawczej*, Prace Filologiczne **LXVI** (2015), 73–83.