Classification in the Speech Recognition

D. Donko, I. Traljic

Abstract—This paper presents specific approach of the classification of the incorrect speech sounds. Sampled data are vowels gathered during the speech therapy with children that have difficulties to pronounce them correctly. Continuous wavelet transformation has been applied on these incorrectly pronounce vowels using Morlet wavelet. Coefficients have been analyzed in the context of three main formants that characterized each of the vowels. The selected coefficients have been classified into main clusters, and have been compared with the one obtained for correct signals. At the end some improvements have been proposed in order to use results in the daily speech therapy and to automate process.

Keywords—Classification, Morlet wavelet, Speech Therapy, Wavelet analysis.

I. INTRODUCTION

SPEECH is the most natural form of human communication and that is the main reason why speech processing has been one of the most explored areas of signal processing [1]. The phonemes of the speech encompass a wide variety of characteristics in time and frequency domain. The two major types of excitation in speech are voiced and unvoiced sounds. Voiced sounds are produced when air flows between the vocal cords and causes them to vibrate. The sound produced by this process is periodic in nature. Voiced sounds in author's native language are A, E, I, O and U. In structure of this language voiced sounds (vowels) are appearing with 41.89% and the most frequent is appearance of vowel 'A', than 'I'', 'E', 'O' and 'U'. This also explains why the correct pronunciation of the vowels is the most important during the speech therapy. Voiced sounds, i.e.. vowels are typically lower in the frequency for longer time duration. High frequency localization gives poor time resolution, and on the other side high time resolution gives poor frequency localization [25]. Wavelet transformations have excellent and deep mathematical properties, making them a well-adapted tool for a wide range of different type of data [3][22][24].

In this paper Morlet wavelet, which is also called 'original' wavelet, has been used for analysis because of its exceptional time-frequency location and its simplicity. Scaling function does not exist for this wavelet and it can be expressed in closed form. For periodic data Morlet wavelet makes clear distinguish between random fluctuation and periodic layers. The biggest disadvantage of this wavelet is that it cannot be used in discrete wavelet transformation and in multiresolution decomposition.

Data that is used in analysis is gathered during the speech therapy with the children of age between 12 and 14 years. Session is supervised by speech therapist that listens and guides patient to pronounce vowels correctly. The process is very time consuming and will benefit from some elements of automation and provision of feedback to patients by the expert system.

Sound features of language can be viewed in several ways. The first way is that the signals are grouped in terms of distribution of acoustic energy in the resonant field, which is caused by speech, which is the acoustic side of the problem, and this approach is given in this paper. Another way is to look at the voice signals from the point of hearing and the ability to receive those signals and to implement parts of the brain, in which case we speak about the perceptual aspects of speech. During the analysis of the speech signal in this paper, we bear in mind that the signals are taken from the people who have speech difficulties, in relation to two basic categories: those with hereditary hearing loss and those with acquired hearing loss, but we considered only the acoustic side of the problem with regard to various factors that led to the inability of proper speech.

Next section describes data processing and analysis of collected data. As the part of signal that we want to analyze further is only non-stationary segment taken on the pitch period, and it is required to determine this period. There are various algorithms to calculate it, and they are based primarily on three approaches: correlation method (time domain) method HPS-Harmonic Product Spectrum (frequency domain) and usage of the wavelet transform (control group). We used algorithm based on the correlation method since it works well for this type of the signals in the time domain.

The third section describes process of classification in the context of three main formants that characterized vowels and propose usage of k-means method. K-means algorithm is one of the most known unsupervised clustering algorithms. Nevertheless, many theoretical improvements for the performance of original algorithms have been explored [9], [10], [11].

Finally, some recommendations have been done and future developments and direction for practical deployment have been proposed.

II. ANALYSIS OF THE SAMPLES

The samples of the speech signals for the children with incapability to correctly pronounce vowels have been taken during the speech therapy. Usually during this therapy, patient

tries to repeat particular vowel after speech therapist's instructions and supervision. One vowel is pronounced constantly, without any accent or emotions, and there is no any other consonant around it. This signal can be considered as the stationary, because there is no change in structure that crate that signal, i.e. there is no change in the vocal tract. It is typical for all vowels that they are almost periodic.

According to the area where they form in the oral cavity, vowels can be divided into:

-front vowels, which are formed by the front of the tongue in front of the mouth (I and E)

-medium vowels, which are formed by slightly raising the body language toward the roof of the mouth (A)

-back vowels, which are formed by the back part of the tongue in the back of the mouth (O and U)

It is well known that the shape of the mouth is the most important part of the vocal tract, and in this paper the analysis boils down to just these elements. The impact of the strength of voice e.g. power is removed by normalizing of all selected signals and reducing energy at the same value. Influence of vibration of vocal cord, or the fundamental frequency F0 is determined at the beginning of the analysis only to extract the signal at the base period. The signal is then interpolated so that all the samples are observed at the same period, so in the final outcome of the research also this impact is removed, i.e. the substantial part the size and shape of the oral cavity in the vocal tract during vowel formation are considered.

For the acoustic field of the vowels it is characteristic that they are formed from the shape of the formant's acoustic energy. Each vowel has three basic formants. Depending of the frequency position of each of the formants, their intensity and their transition, we obtain the acoustic characteristics of vowels. The lowest limit for the first formant is 100 Hz, and the highest for the third formant is 3000 Hz. For the female voices these limits are increased by 18% and for children by 25%. The length of the vocal tract is usually 17 centimeters, but it may vary due to lifting and lowering the larynx, and changing the shape of lips. Fig.1 shows frequency position for all three formants for analyzed vowels, and their down and up limits.

It is common to characterize formant according to their position on the frequency scale. The lowest formant is called F1, and its lower limit can be set around 100 Hz. What is the position of this formant lower, the tongue is closer to the roof of the mouth. The following formant is a formant F2, and its position as the other formant depends on the spoken vocals. The value of the formant peak is proportional to the position of the tongue, and as it is more prominent the frequency is higher. This formant is also significantly affected by the shape of the lips. Rounded lips will cause that the formant is much lower in frequency scale. The upper limit of the third formant F3 is about 3000 Hz, and its position also has a major impact on the quality of the spoken vowel. Variations of formants' positions depend not only on the personalities who speaks, but also from the influence of voices surrounding the vocals, as well as accent, tone and duration of the observed vowel.

Analysing the table I we can notice following [20]:

-The vowel "A" includes the first formant frequency band of 281 Hz. Distance between frequencies peaks for the first two formant is much lower than for the vowel "E" and "I". In the case of possible upper limit of the first formant and the lower limit of the second formant it can come to the critical convergence, but the formants still retain their specificity. The distance between the second and third formant is sufficiently strong. For vowel "A" it can be said that formants' position a distributed in the way that they characterized this vocal with their anti-formants properties.

-For the vowel "E" the first formant range comprises of 219 Hz. Distance between the peaks of the first and second formant and second and third formant is very clear. Distribution of the formants in the structure of this vowel is expressed enough so it cannot come in any case to their overlaps. This means that the definition of its acoustic structure is clear enough so that the normal ear can easily determine its acoustic image.

-The vowel "I" has the first formant range of 218 Hz. The distance between the top of the first and second formant, i.e. the first anti-formant, is big enough and is 2064 Hz. The distance between the second and third formant, i.e. the second anti-formant, is much smaller but still sufficient to avoid overlapping. The vocal is clearly audible defined and the basic acoustic elements are sufficiently separated from each other.

-The vowel "O" has the frequency band for the first formant of 235 Hz. The smallest possible distance between two first formants is only 84 Hz, so we have a case similar to the vowel "A" that the first formant almost overlaps. The smallest possible difference between the second and third formant is 538 Hz, which is enough to keep the formant features of the pronunciation of vowels, even if the accent is present.

-The vowel "U" has the frequency band for the first formant of 208 Hz. If the upper limit of the first formant raise the value of standard deviation, which is 79 Hz, a lower limit of the second formant down the value of standard deviation, which is 74 Hz for the second formant, there would be overlap in these two formant. For the second and third formant distance is sufficient so it cannot come to overlapping.

It can be seen that within each vowel formants are distributed in such way that it cannot come to overlapping, except in the case of the first anti-formant for vowel "U". However, it is interesting to see what is happening with the formants overlapping between different vowels. It is easy to see on Fig. 1 where all formants are presented together with their formants position. Different relation can be observes at the frequency scale for each vowel and for their mutual relations.
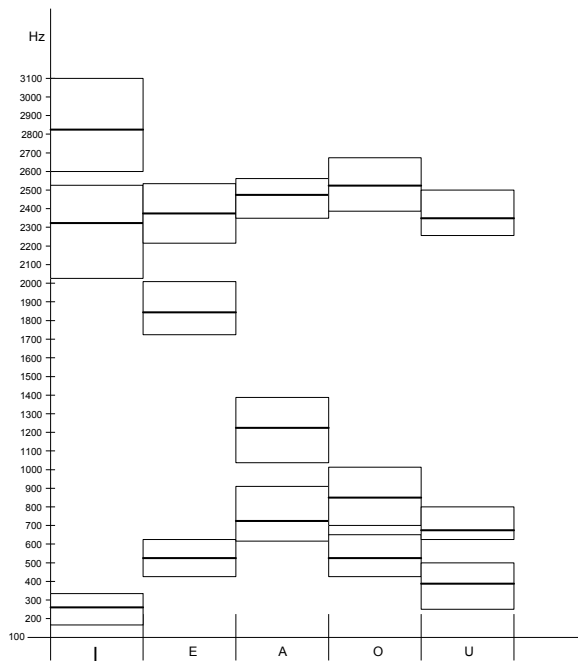
Fig. 1 position of the three main formants for each vowel

The shape of the mouth is the most important part of the vocal tract, and this paper comes down to just these elements. The influence of voice volume i.e. energy is disregarded by normalizing all incoming signals and reducing energy at the same value. Influence of vibration of vocal cord or the fundamental frequency F0 is determined at the beginning of the analysis only to extract the signal on the basic period. The signal is then interpolated so that all samples are observed at the same period, so that the final results this influence is also eliminated, and a substantial part of it takes in consideration the size and shape of the oral cavity in the vocal tract during vowel formation.

Let's take a look at the vocal tract in the formation of vowel "I". The lips occupy a neutral position or easily stretched. Position of the tongue is not constant but depends on the sounds that surround it, but nevertheless has its articulatory position. The space between the front of the tongue and the inside of the upper incisors is very small, but sufficient to avoid friction. In addition, the lower jaw is lifted to raise the tongue to the roof of the mouth. Synchronization of the movement of the lower jaw, tongue and neutral position of lips creates a unique dynamic-articulatory system for creating vocal I.

It is usual that formant is characterized by its position on the frequency scale. What is the position of the lower formant lower, the tongue is closer to the roof of the mouth. The value of the second formant position is proportional to the position of the tongue tip, and more advanced in front it is, the higher frequency is. This formant is also significantly affected by the shape of the lips. Rounded lips cause that the formant is significantly lower in the frequency scale. The position of the highest third formant is influenced by accent, tone and vowel duration.

There are many ways to analyze end extracting different parameters of the voice signal [23]. Some authors propose a method for epoch extraction which does not depend critically on characteristics of the time-varying vocal-tract system and exploits the nature of impulse-like excitation using zero resonance frequency filter output [6]. In some research also pitch period is first estimated for short time speech signal, and then the next period data are predicted by the previous one with one parameter according to the least-mean-square-error criterion [7]. For the further analysis the signal has been considered only on the pitch period. Pitch period corresponds to the frequency F0.

To discover that period, autocorrelation function has been used. Mean value of the distances between maximums of the function is calculated.

Table I. Formants and their frequency characteristics

Vowel A			
	1. Formant	2. Formant	3. Formant
Low end (Hz)	624	1080	2345
Peak (Hz)	763	1245	2464
High end(Hz)	905	1390	2576
Vowel E			
	1. Formant	2. Formant	3. Formant
Low end (Hz)	446	1759	2241
Peak (Hz)	545	1894	2395
High end(Hz)	665	2036	2566
Vowel I			
	1. Formant	2. Formant	3. Formant
Low end (Hz)	177	2040	2590
Peak (Hz)	261	2325	2868
High end(Hz)	335	2523	3130
Vowel O			
	1. Formant	2. Formant	3. Formant
Low end (Hz)	430	780	2382
Peak (Hz)	553	884	2484
High end(Hz)	665	981	2626
Vowel U			
	1. Formant	2. Formant	3. Formant
Low end (Hz)	279	632	2299
Peak (Hz)	380	746	2378
High end(Hz)	488	814	2470

Once signal has been taken on the pitch period it is non-stationary signal, and wavelet analysis is becoming very powerful tool for the further analysis. Two important modifications and adjustments of this signal have been done to avoid taking into consideration signal parameters that are not relevant for this analysis:

-Each signal has been rescaled to the same period, and it is 600 samples for each signal, which is for this signals equal to 2,33 ms. This has to be done, because it will be irrelevant value of the laryngeal tone F0, which can vary from person to person

-Each signal is normalized that energy is equal 1. In that case higher value of the coefficient is not result of the loudness of the pronounced vowel.
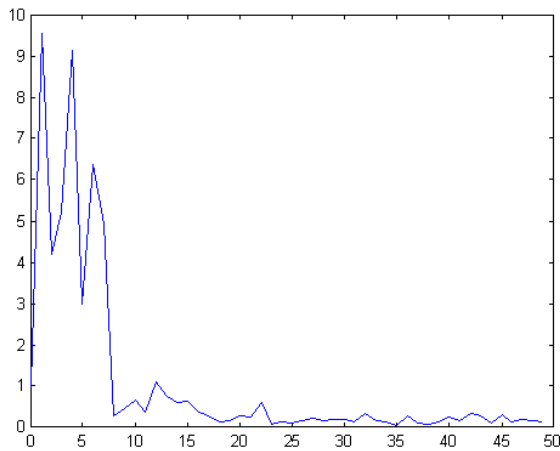
Fig. 2 Fourier transformation of the correctly pronounced vowel "A"
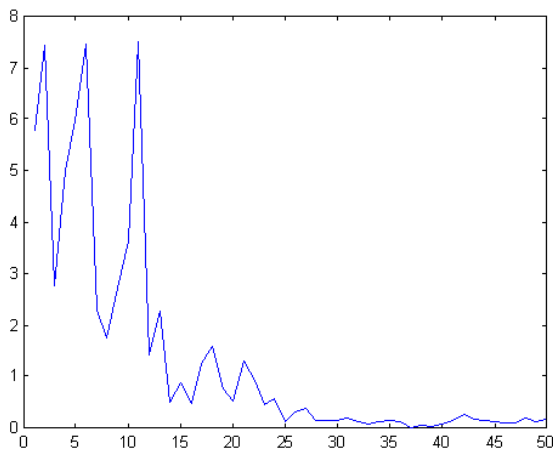

Fig. 3 Fourier transformation of the incorrectly pronounced vowel "A"

Observed sample interval for selected vowel is still kept to 2.33 ms. Before we move on wavelet analysis of signals it is interesting to take a quick review of the Fourier transformation of non-stationary signals, i.e. what are the frequency components present in the signal. Since we are looking into formant frequency spectrum, we have three formants that have to be presented at certain frequencies, according to Table I. These frequency peaks are 763Hz, 1245Hz and 2646Hz. Frequency spectrum has the appearance as shown in Fig. 2. Range of [0.300] sample correspond to the frequencies [0, 100 kHz]. Fig. 2. presents the spectrum from 0 to 16.5 kHz, and we can observer peaks in the points 3, 4.8 and 7.6, i.e. the frequencies of 1003Hz, 1600Hz and 2541Hz. Formant positions are slightly shifted towards higher frequencies, which is expected considering that this is a sample of vowel pronounced by children. The Fig. 3 presents the Fourier transform of incorrectly pronounced vowels, and here we can see that usage of Fourier transformation does not give good results for identifying faulty vowels, because the positions of the formants are quite close, and the analysis itself does not

lead to some important conclusions.

We will also have for the other vowels three peaks that will correspond to the formants, and what distinguishes them are their positions in the frequency spectrum. For the people with difficulties to speak, exactly the lack or incorrect position of a formant causes that the voices are not sufficiently understood. However, the main disadvantage of basic frequency characteristics is that we do not have any information about the time, i.e. the information about the frequency components is not telling us anything about the position of these spectral components in time. This can be overcome by using the wavelet transformation. Values of these formants are small compared to the frequency of F0, and they the most important for the voice recognition, which is precisely the weakness of the Fourier transformation.

III. SIGNALS CLASSIFICATION

Continuous wavelet transformation transforms a continuous signal into a new form that is a function of continuous shift and continuous scale. The coefficients are function of the scale and position, e.g. we can say that they are function of frequency and time, and are calculated as:

$$C(s,\tau) = s^{-\frac{1}{2}} \int_{-\infty}^{+\infty} f(t)\psi(\frac{t-\tau}{s})dt$$

In this paper continuous wavelet transformation has been applied to the signals using Morlet wavelet. Morlet wavelet gives some advantages compare to the other wavelet families, such as exceptional time-frequency location, faster algorithm and fewer problems with edge affects. The Fourier spectrum of complex Morlet wavelet is real and the complex Morlet wavelet does not affect the phase of a signal in complex domain. That result in a desirable ability to very accurate detects the singularity characteristic of a signal. This wavelet has been successfully used in the signal de-noising for bearing fault diagnosis [8] [14] [15]. Also many other different algorithms have been developed based on this wavelet. New symbol rate estimation algorithm, based on Morlet wavelet transform and autocorrelation has been proposed [13]. Iris recognition method based on the coefficients of Morlet wavelet transform has been also developed [16][17]. One of these approaches makes normalization to the iris image, makes one dimension Morlet wavelet and gives the recognition results with recognition rates up to 99.946%. Morlet wavelet has been also used for different power-quality detection systems [18][19].

Table II gives position of each formant taking into consideration time and scale. Analyzing data down and up limits for formants can be also determined in similar way.

Table II. Position of each formant for correct vowel from Fig.4

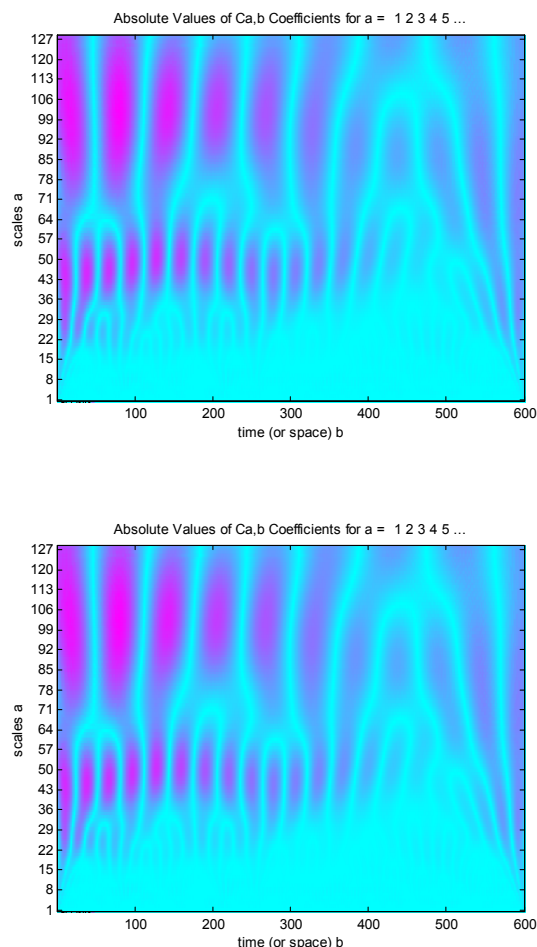| Time	Scale	Formant
16, 36, ...	22	3rd formant
4,42,68,...	46	2nd formant
24, 83, 136,...	106	1st formant

Fig. 4 continuous wavelet transformation of correct and incorrect vowel A using 'Morlet' wavelet

Fig. 4 give an overview of the correct and incorrect vowel A. We can see significant differences in the displayed coefficients. At the correctly pronounced vowel there is almost no defined boundary between first and second formant and they are gradually complemented each other, while in the case of incorrect vowel that border is significant. The position of the first formant is generally the same for both samples, which is normal considering fact that people with hearing impairment have fewer problems with low frequencies. Here we should mention another remarkable feature of such an analysis of the speech signal. Namely, we can see that the positions of the maximum value of coefficients for each scale, if we look to the formant side, are not having always the same value. This is

expected because the samples were taken from different people and there is always a tolerance for the formant peak. However, what builds the voice and makes it understandable is just possible overlapping of the formants, and allowed distance between them. This view of the coefficients gives exactly this pattern, the observation that for different patterns can be easily seen as very useful for diagnostic purposes. Other details that need attention is the pattern of placing highest coefficients considered in the relation to time i.e. scale shifts of mother wavelet. This scale view gives us very indicative information about the components.

The areas with high intensity i.e. with high coefficients value have been considered only, so huge reduction has been performed using threshold elimination and interpolation.
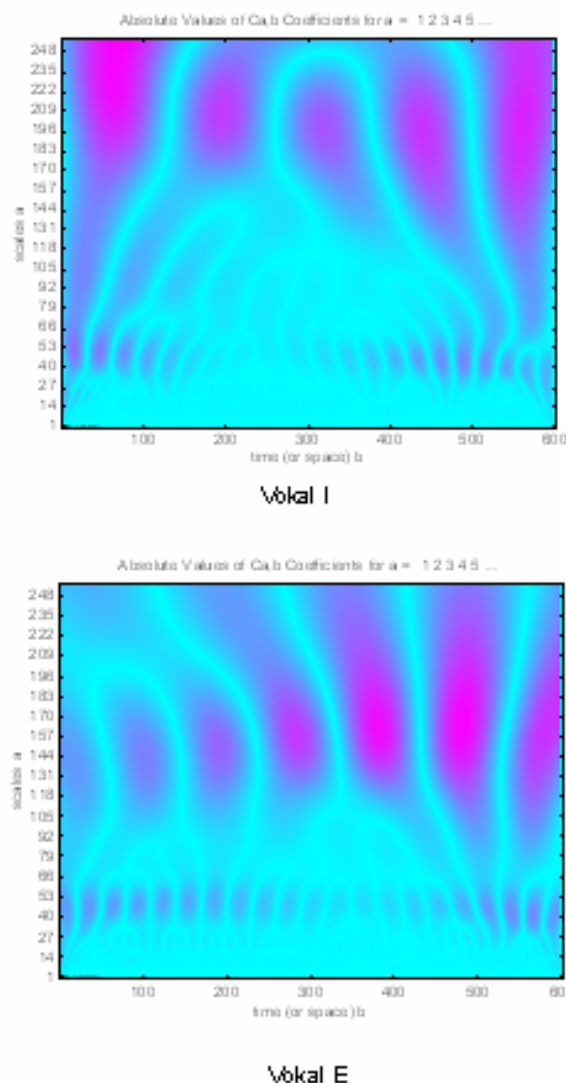

Vokal I

Vokal E

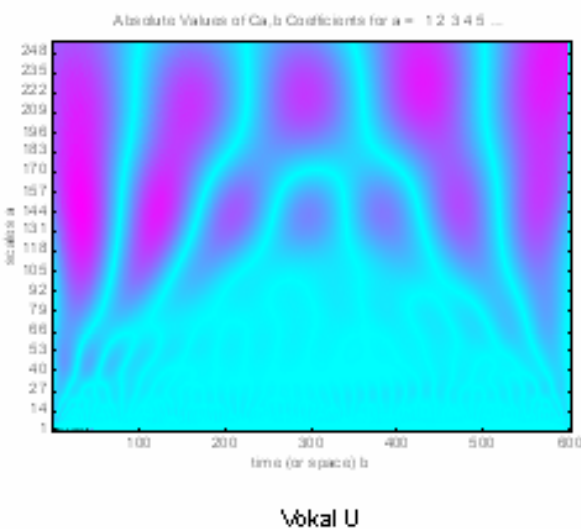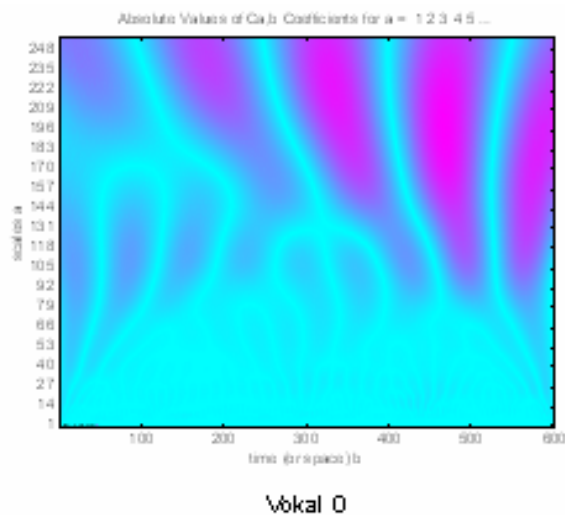Fig. 5 continuous wavelet transformation of vowels I and E using Morlet wavelet

Fig. 7. incorrect vowel "I"

Fig. 6. continuous wavelet transformation of vowels O and U using Morlet wavelet

Fig. 8. incorrect vowel "O"

If we take a look at the value of the coefficients for the front vowels I and E, it can be observed the presence of the third formant at the scale around 40, and the strong presence of the first formant, while the intensity of the second formant is very low. In addition, the coefficients for the first formant are at a slightly lower value of the scale, which corresponds to a higher frequency as it is shown on Fig. 5. Back vowels have almost negligible value ratios at low scale values, i.e. high frequencies corresponding to the third formant. In the analysis of these vowels, it is obvious that the front vowels have a lower resonance field in the oral cavity due to the participation of the front of tongue in their creation. This basic difference in their creation can be associated with this distribution of the coefficients. For the vocal A we can observe three levels of the scale with the high values of the coefficients, but they no longer appear as evenly as for the front vowel. Also, the third formant is located at a slightly higher scale, corresponding to a lower frequency than the front vowel E.
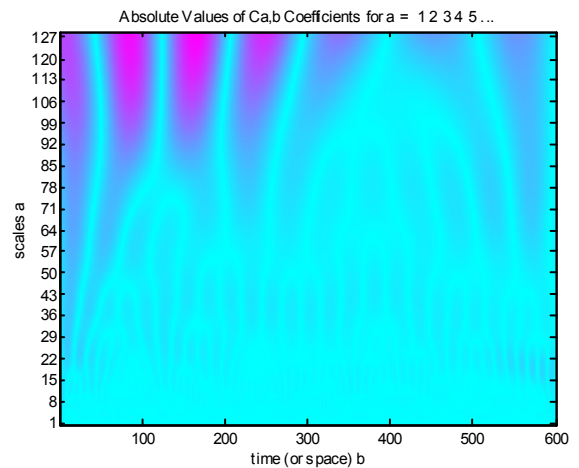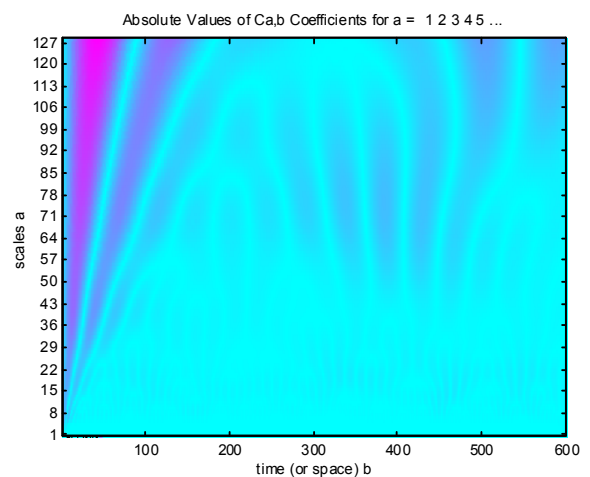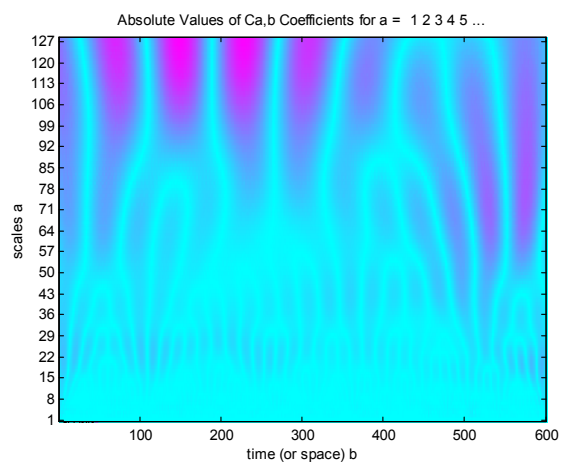
Fig. 9 incorrect vowel "E"

In Fig. 9 we can see that the incorrect vowel "E" comes to

the uneven distribution of the second formant, and intensity of these coefficients is much smaller than for the correct vowel.

Similar observation we can apply for the vowel "I", and considering that it is a front vocals, a significant impact is in the front position of the top of tongue, which incorrect position can greatly affect the values of the coefficients.

During the research it is also discovered that the biggest differences for correct and incorrect vowel "O" and "U" are happening at the top of the scale i.e. for the lowest frequencies. It is to be expected considering the fact that for these vowels the vocal tract occupies a much larger cavity and the impact has withdrawn the last of the tongue in the oral cavity. For vocal in these changes are very small and are expressed only at the end of the timeline.

We proposed use of k-means method for classification and for diagnostic purposes. It is a popular clustering method used in many different fields of computer science, such as data mining, machine learning and information retrieval. This algorithm takes as input parameter number of cluster, and partitions set of n objects into k clusters. This produces high intra-cluster similarity but low inter-cluster similarity. The method is relatively scalable and efficient in processing large data sets Proposed clustering method takes into consideration three clusters matching to three formants taking most of the signal scale characteristics. Thanks to localization characteristic of wavelets, it is possible to identify position in time when these formants are presented.
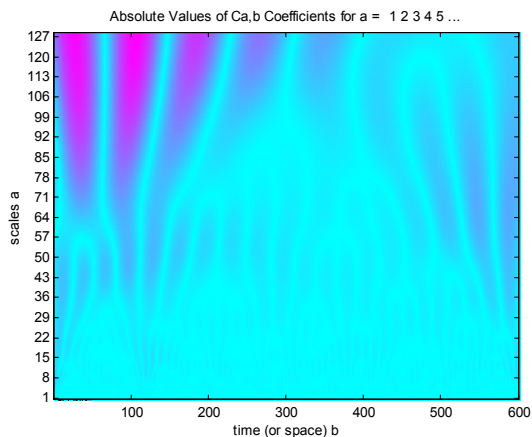

Fig. 10. incorrect vowel "U"

Referent model used for classification are data gathered into three clusters with position matching the scale of correct formants. Number of clusters for irregular vowels is unchanged, but algorithm is forced to position them into previous space. Any difference from correct reference will indicate anomalies and changes in the voice characteristics. E.g. if the position of the cluster that reference to first formant is lower than expected, speaker tongue is too close to the roof of the mouth. Clustering can be part of integral intelligent system which can lead patient to make correction providing right feedback and having expected result without supervision of therapist.

IV. CONCLUSION

Our experimental results show that there is strong link between formant of vowels and decomposition of signal using wavelets. It is shown that the most important role in the formation of vowel change is in the shape of the vocal tract. We described distinct differences in the front and back vowels, and causes of such differences, as well as the equivalents of such changes in the wavelet coefficients. Applying mechanisms of classification, in our case using k-means method, it is possible to establish strong relation between formants' position and incorrect pronunciation. Major objective and future work based on results presented here is development of expert system that will use feedback, and provide guidelines to the patients during therapy. The exact match of clusters and differences toward basic model will provide some basic tips from the expert system. So feedback with corrective steps can be used to guide patients through the therapy.

REFERENCES

[1] B. H. Juang Ed., "The past, present, and future of speech processing", *IEEE Signal Processing Magazine*, 24-48, May 1998.
[2] Louis R. Litwin, Jr., "Speech coding with wavelets", *IEEE Potentials*, 38-41, April/May 1998.
[3] A. Neocleous and P. A. Naylor, "Voice source parameters for speaker verification", *Proc. Eur. Signal Process. Conf.*, p.697 , 1998.
[4] I. Daubechies, "Recent results in wavelet applications", *Journal of Electronic Imaging, 7 (no. 4)*, pp. 719-724, 1998.
[5] Wagsta_, K., Cardie, C., "Clustering with instance-level constraints", *Proceedings of the Seventeenth International Conference on Machine Learning,* Palo Alto, 2000
[6] Murty, K.S.R., Yegnanarayana, B., "Epoch Extraction From Speech Signals", *IEEE Transactions on Audio, Speech, and Language Processing,;* Nov. 2008 Volume: 16 Issue:8; 1602 – 1613
[7] You Chunyan , Bai Sen, "Robust Information Hiding in Speech Signal Based on Pitch Period Prediction", *Computational and Information Sciences (ICCIS),* 2010, 17-19 Dec. 2010
[8] Zhou, J.-H., Yang, X, "Reinforced Morlet wavelet transform for bearing fault diagnosis", *IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society*, 7-10 Nov. 2010
[9] Bai Hong-Tao, "K-Means on Commodity GPUs with CUDA", *2009 WRI World Congress on Computer Science and Information Engineering*, April 2009
[10] Isa, N.A.M., " Adaptive fuzzy moving K-means clustering algorithm for image segmentation", *IEEE Transactions on Consumer Electronics*, Volume: 55 Issue:4; November 2009
[11] Sulaiman, S.N., Isa, N.A.M, "Denoising-based clustering algorithms for segmentation of low level salt-and-pepper noise-corrupted images", IEEE Transactions on Consumer Electronics Volume: 56 Issue:4; November 2010
[12] Deepak S. Turaga, "On K-Means Cluster Preservation Using Quantization Schemes", *2009 Ninth IEEE International Conference on Data Mining,* December 2009
[13] Yong Gao, "A symbol rate estimation algorithm based on Morlet wavelet transform and autocorrelation", *IEEE Youth Conference on Information, Computing and Telecommunication, 2009*. YC-ICT '09. September 2009
[14] Zhou, J.-H. and Yang, X., "Reinforced Morlet wavelet transform for bearing fault diagnosis", *IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society*, November 2010
[15] Hui Li, "Complex Morlet wavelet amplitude and phase map based bearing fault diagnosis", *8th World Congress on Intelligent Control and Automation (WCICA)*, 2010, July 2010

[16] Lin Zhonghua and Lu Bibo, "Iris Recognition Method Based on the Coefficients of Morlet Wavelet Transform", *International Conference on Intelligent Computation Technology and Automation (ICICTA),* May 2010

[17] Hariprasath, S. and Mohan, V., "Biometric personal identification based on iris pattern recognition using Wavelet Packet Transform", *International Conference on Computing Communication and Networking Technologies (ICCCNT),* July 2010

[18] Chia-Hung Lin and Chia-Hao Wang, "Adaptive wavelet networks for power-quality detection and discrimination in a power system", *IEEE Transactions on Power Delivery,* July 2006

[19] Tse, N.C.F., "Practical application of wavelet to power quality analysis", *IEEE Power Engineering Society General Meeting,* October 2006

[20] Kostić Đorđe "Metodika izgradnje govora u dece oštećena sluha" "Methods of construction of talk in children with impaired hearing", Savez društava defektologa Jugoslavije, 1982

[21] S. Dogan, B. Barac, "Govor i poremećaji govora"," Speech and speech disorders" Sveučilište u Zagrebu, Visoka defektološka Škola, Zagreb 1971

[22] 1.Wael Al-Sawalmeh, Khaled Daqrouq, Abdel-Rahman Al-Qawasmi, Tareq Abu Hilal, "The Use of Wavelets in Speaker Feature Tracking Identification System Using Neural Network", *WSEAS TRANSACTIONS on SIGNAL PROCESSING,* Issue 5, Volume 5, May 2009. ISSN: 1790-5052, pages 167-177.

[23] Jalal Karam, "A comprehensive approach for speech related multimedia applications", *WSEAS Transactions on Signal Processing , Volume 6 Issue 1,* January 2010

[24] Wael Al-Sawalmeh, Khaled Daqrouq, Abdel-Rahman Al-Qawasmi, Tareq Abu Hilal, "The use of wavelets in speaker feature tracking identification system using neural network", *WSEAS Transactions on Signal Processing , Volume 5 Issue 5,* May 2009

[25] Andrej Žgank, Tomaž Rotovnik, Mirjam Sepesy Maučec, "Slovenian spontaneous speech recognition and acoustic modeling of filled pauses and onomatopoeias", *WSEAS Transactions on Signal Processing , Volume 4 Issue 7,* July 2008