FACTA UNIVERSITATIS (NIŠ)
SER.: ELEC. ENERG. vol. 23, no. 3, December 2010, 379-393

Recognition of Vowels in Continuous Speech by Using Formants

Biljana Prica and Siniša Ilić

Abstract: Speech consists of acoustic pressure waves created by the voluntary movements of anatomical structures in the human speech production system. These waveforms are broadly classified into voiced and unvoiced speech. Voiced sounds (vowels for example), produce quasi-periodic pulses of air which are acoustically filtered as they propagate through the vocal tract. The main distinction between vowels and consonants is that vowels resonate in the throat. Formants are exactly the resonant frequencies of a vocal tract when pronouncing a vowel. In this paper we attempt to carry out Vowel Recognition through Formant Analysis in Serbian language, wherein we detect which of the five Serbian vowels is spoken by the Speaker. Here we describe a standard approach for classification of vowels in continuous speech based on three formants: F1, F2 and F3. We have investigated the correlations between formants in each vowel and developed the algorithm to reduce the overlap of different vowels in F1-F2 and F2-F3 planes.

Keywords: Serbian speech, recognition of vowels, continuous speech, formants.

1 Introduction

ANALYSIS and presentation of the speech signal in the frequency domain are of the great importance in studying the nature of speech signal and its acoustic properties. The prominent part of speech signal spectrum belongs to formants that correspond to the vocal tract resonant frequencies. The quality of some of the most important systems for speech recognition and speech identification as well as systems for formant based speech synthesis are dependent on how accurate the formant frequencies are determined.

Manuscript received on April 22, 2010.

The authors are with Faculty of technical sciences in Kosovska Mitrovica, Kneza Miloša 8, 38220 Kosovska Mitrovica Serbia (e-mails: kaca._pepi@nadlanu.com and sinisa.ilic@pr.ac.rs).

Serbian language consists of 30 phonemes, of which 5 vowels and 25 consonants. Vowels are /**a**/, /**e**/, /**i**/, /**o**/ and /**u**/. Although there are only 5 vowels, they appear in the Serbian language in 44.6% of the total occurrence of phonemes. Together with nasals (m, n, nj) and semi-vowels (v, j) there are about 60% of phonemes that have formant structure compared to the total occurrence of all phonemes [1].

The combinations of two phonemes in syllables found in the Serbian language are: the consonant-vowel (CV) and vowel-consonant (VC) in 40.685% of cases. If we consider occurrences of vowels (V), already mentioned two phonemes syllables (CV and VC) and syllables that consist of three phonemes, such as: consonant-consonant-vowel (CCV) and consonant-vowel-consonant (CVC), then it amounts to 92.5% of all syllables present in the Serbian language. In 25% vowels can be found at the beginning of words.

The statistics presented shows how recognition of vowels plays an important role in recognition of continuous speech in the Serbian language. This paper describes the recognition of the Serbian language vowels by using the formant analysis.

1.1 Formants and formant structure of Vowels in Serbian speech

Basic acoustic properties of vowels can be seen in their short-time spectra [2]. The spectrums of 5 vowels in Serbian speech obtained by using of 19 filters bank in the time frame of several tens of seconds are shown at Figure 1. Central frequencies and bandwidths of mentioned filters are selected according to the Holms' filters bank [3] and presented at Table 1.

Table 1. Holms' filter bank

filter	1	2	3	4	5	6	7	8	9	10
Central freq. [Hz]	240	360	480	600	720	840	1000	1150	1300	1450
Bandwidth	120	120	120	120	120	120	150	150	150	150

filter	11	12	13	14	15	16	17	18	19
Central freq. [Hz]	1600	1800	2000	2200	2400	2700	3000	3300	3700
Bandwidth	150	200	200	200	200	300	300	300	300

As can be seen from Figure 1, all spectra have harmonic structure. The peaks that exist at the output of filters with the lowest sequence number (that correspond to lowest frequencies) represent fundamental frequency of a speaker. Other peaks correspond to the resonant frequencies of the vocal tract - ie. formants. Those frequencies are the frequencies wherein the concentration of acoustic power is the

largest. The spectrum of phonems can consist of several formants, but the first three are most important for recognition. Formants are present not only at vowels, but recognition of the vowels based on them is easier and gives better results.

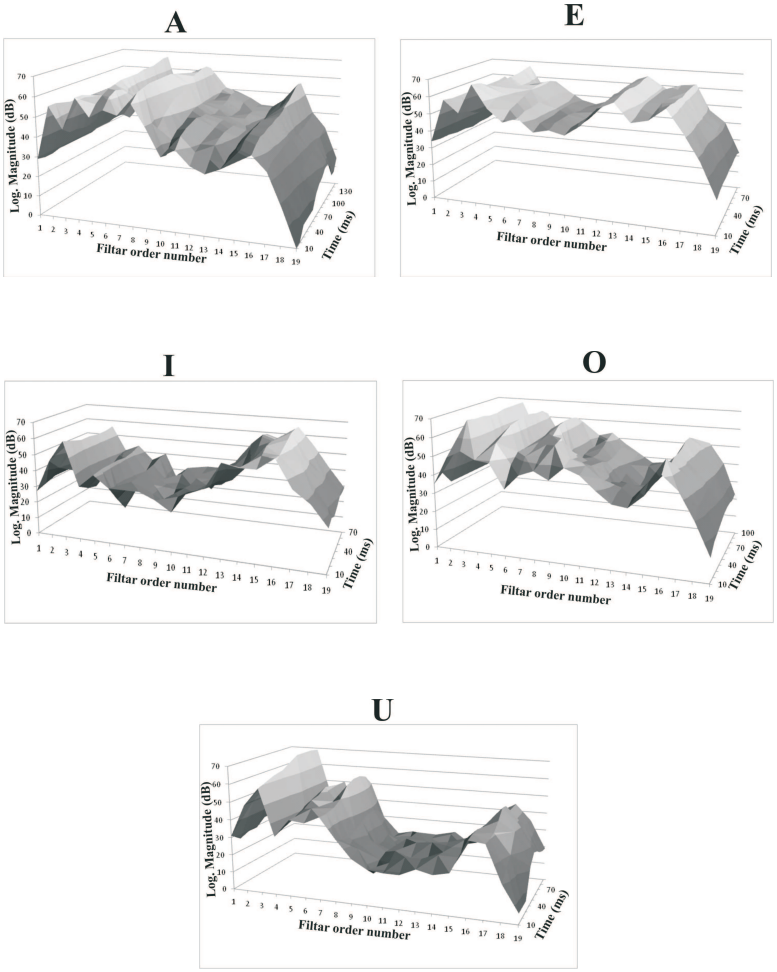


Fig. 1. Short time spectrums of vowels

During the vowels pronunciation, the frequency of the first formant (F1) can be found in range from 250Hz to 1000Hz. The tongue is closer to the hard palate the frequency of F1 is lower. The frequency of the second formant (F2) can vary from 550 Hz up to 2700 Hz and it depends on front and back position of the tongue. The lower frequency for F2 can be achieved by rounding the lips [4]. Third formant (F3) is important for quality and clarity of pronounced phoneme.

In this paper we describe the recognition of vowels based on position of for-

mants on frequency axis in continuous speech. It is well known that positions of formants in vowels depend also on co-articulation with other phonemes in continuous speech, and it makes recognition harder.

2 Current Research in Continuous Speech Recognition

The recognition of vowels by using formant frequencies in Arabic speech is described in [5] and [6]. In [5] authors have described research in segmentation and identification of Arabic vowels in continuous speech based on transitions in formant frequencies. They have developed the recognition system with accuracy up to 90% from the speech with 1000 vowels. Alotaibi et al. [6] have researched Arabic vowels by using their characteristics in time and frequency domain, and by using formant frequencies. They have developed the recognition system based on HMM and determined experimentally the frequencies of the first and second formant (F1 and F2). The research has been performed on signal obtained by isolated spoken words, by focusing on centers of vowels in time frame in order to avoid co-articulation. The mean value of results obtained for recognition was about 91.6%.

Yusof et al. [7] have presented a new method for vowel recognition by using Autoregressive Models (AR). They have used syllables: “KA, KE, KI, KO, KU” in order to present appropriate vowels and have obtained excellent results of 99%, but they also have got a high percentage of wrongly recognized vowels.

In recent similar studies of English vowels, Kodandaramaiah et al [8] have described the standard approach for the classification of vowels based on formants. They have got 80% to 95% of accuracy in speaker recognition based on Euclidean distance.

Kocharov [9] has developed a system of recognition of vowels in the Russian language which is based on synchronization with the pitch period. The central phase of the system is segmentation of characteristics. There is achieved the recognition of 87.70% for isolated vowels and 83.93% for the vowels within a word.

3 Using of LPC Method in Speech Analysis

Linear Predictive Coding (LPC) is a tool used mostly in audio signal processing and speech processing for representing the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive model. It is one of the most powerful speech analysis techniques, and one of the most useful methods for encoding good quality speech at a low bit rate and provides extremely accurate estimates of speech parameters (taken from Wikipedia). We have used LPC to determine coefficients of recursive filter with all poles [10] in

order to obtain envelope of transfer function of vocal tract. The scheme of such filter is shown at Figure 2.

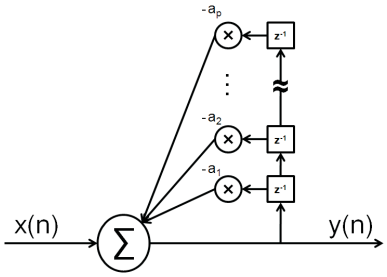


Fig. 2. Scheme of recursive LPC filter of order p

The transfer function of the filter is:

$$H(z) = -\sum_{k=1}^{p} a_k z^{-k} \tag{1}$$

When coefficients of filter are determined, the signal samples at the output of the filter represent the optimal prediction of future samples based on current speech samples. The coefficients of filter are determined by minimizing the squared error between the real samples $s(n)$ and predicted samples at the output of the filter $\tilde{s}(n)$. The estimated sample $\tilde{s}(n)$ depends on previous p samples according to:

$$\tilde{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{2}$$

where a_k are coefficients of LPC filter.

The difference between real and estimated sample is prediction error and it is represented by formula:

$$e(n) = s(n) - \tilde{s}(n) \tag{3}$$

If we assume that the coefficients of LPC filter are constant within some time frame, we can determine coefficients' values by minimizing the square of prediction error E:

$$E = \sum_{n} [e(n)]^2 \tag{4}$$

At the Figure 3 is shown the simplified model of speech generation at human body, where the vocal tract is represented by LPC filter. In this simplified model, the influence of glottal pulses shape, vocal tract transfer function and radiation on

the lips are represented by time variable digital filter whose transfer function $T(z)$ has the form:

$$T(z) = \frac{1}{1-H(z)} = \frac{1}{1-\sum_{k=1}^{p} a_k z^{-k}} \tag{5}$$

The excitation of the filter can be done by pulse train (for voiced phonemes) or random noise generator (for un-voiced phonemes). So the parameters for system presented are: switch position (to select generator type), period of fundamental frequency (for pulse train) and filter coefficients a_k. As we are interested in generation of vowels, that are voiced phonemes, and as the pulse train has flat spectrum, it is clear that the filter shapes the spectrum of spoken vowels.

The filter with all poles is a good approximation of the vocal tract transfer function for voiced, non-nasal phonemes. In the case of modeling nasal and voiced fricative sounds transfer function of model must have both - zeros and poles.

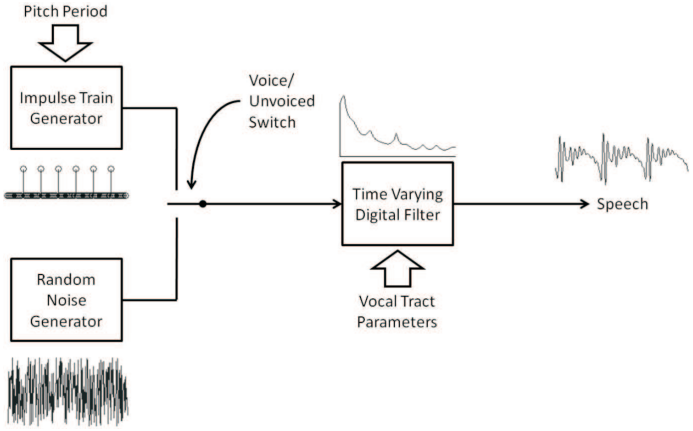


Fig. 3. Simplified model of speech generation

3.1 SFS application

It was decided to perform the speech signal processing using SFS package because of many reasons. SFS package is the property (copyright) of University College London, but is currently available free of charge to organizations that deal with non-profitable research. One of the available applications of SFS package is **fmanal** that is based on linear prediction method and gives formant frequencies and their amplitudes. **Fmanal** uses linear prediction technique to the windowed sequences of speech signal and solves predictor's polynomial gaining values of spectral peaks

- formants. Width of the window can be constant with defined overlapping, or can be adapted to the value of pitch period of voiced sequences of speech signal. Pre-emphasis, Hamming window function and autocorrelation method is used for all analysis. In our analysis we have used Hamming window of constant width of 20 ms length with 10 ms overlap.

The output from the SFS suite is the matrix with the coefficients shown at Table 2.

Table 2. The output of application fmanal from SFS package.

t	ANOT	F1	A1	F2	A2	F3	A3	F0
10		1151	39	2059	55	2081	56	0
20		744	41	1552	39	2028	54	0
				. . .				
620	B	804	71	2280	84	2962	76	186
630	B	408	89	2166	75	2453	79	212
640	I	434	86	2303	80	2626	85	217
650	I	442	83	1941	63	2440	85	220
660	I	440	82	1898	60	2407	87	219
				. . .				
1300	A	635	97	1606	87	1950	76	195
1310	A	698	95	1642	88	2666	82	195
1320	A	764	98	1533	92	2055	75	193
1330	A	774	100	1481	93	1896	79	193

The column t represents time in ms, ANOT is manually added letter by user, F_n and A_n ($n = 1,2,3$) are frequencies and amplitudes of appropriate formants respectively and F_0 is fundamental frequency.

Each row in Table 2 represents characteristics of Formants and fundamental frequency of the speech signal within the time frame of 20ms. We have used the term **sequence** in the ongoing text to denote sequence of rows in this table.

4 Method for Analysis of Formants

We have analyzed the formant frequencies of first three formants of Serbian vowels simultaneously. The areas that vowels occupied in F1-F2-F3 coordinates are shown at Figure 4 and Figure 5.

The discrimination of vowels based on first two formants is not possible because of big overlapping in F1-F2 plane (see Fig. 5a). But if third formant is joined to decision making, the overlapping is decreased significantly.

Based on the averaged values we have got from the matrix shown at Table 2, we have determined the histograms of distribution of formant's frequencies F1, F2

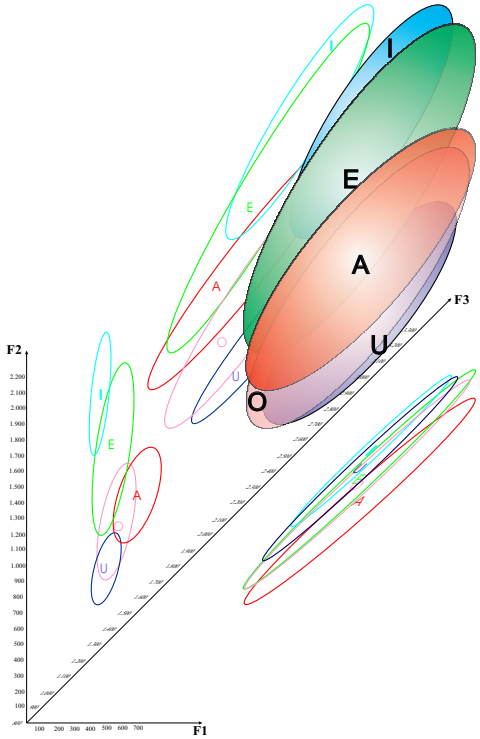
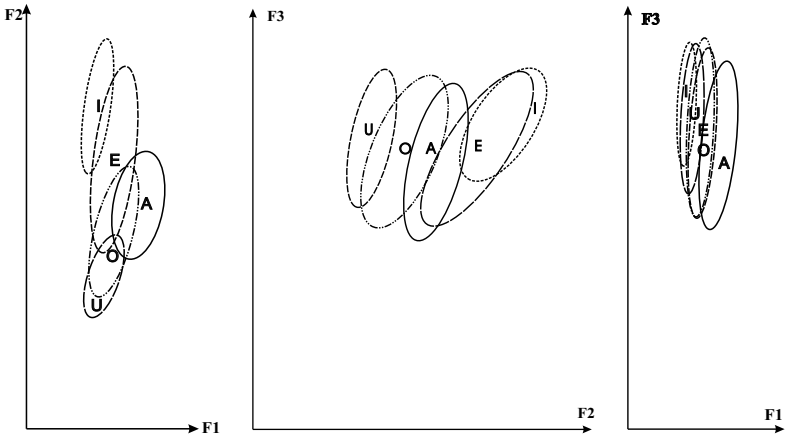


Fig. 4. Frequencies of the first three formants of Serbian vowels



(a) F1-F2 (b) F2-F3 (c) F1-F3

Fig. 5. Frequencies of the first three formants of Serbian vowels projected at appropriate planes

and F3 for Serbian vowels and presented them at Fig. 6.

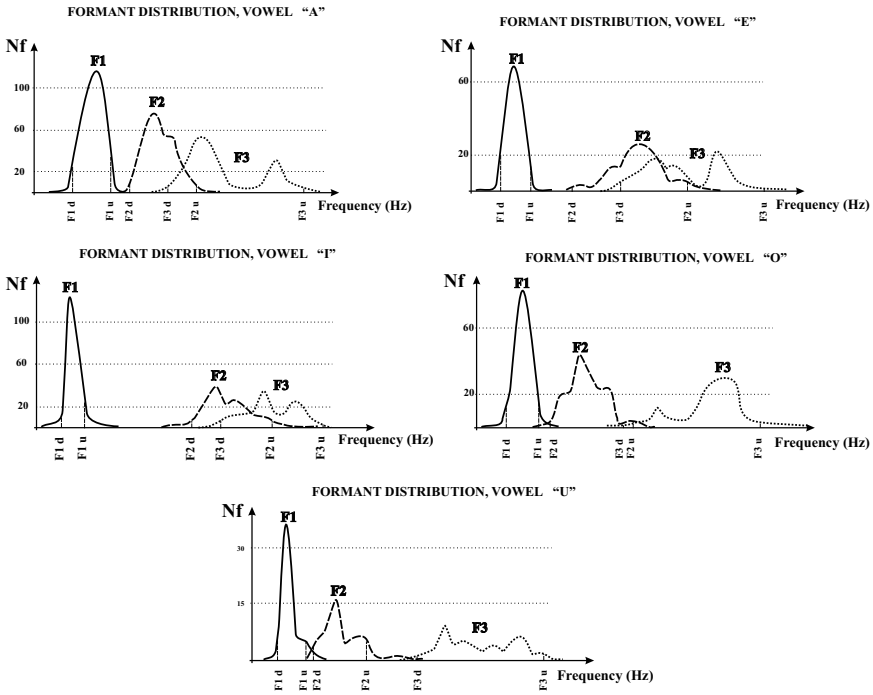


Fig. 6. Histograms of formant frequencies of Serbian vowels

By analysis of histograms, we have defined lower (Fn_d) and upper (Fn_u) bounds for all formants in the way that 90% of values in histograms belong to defined bounds. Also, we have defined the boundaries of the areas that vowels occupy in F1-F2, F2-F3 and F1-F3 planes. The example is shown at Figure 7 for F1-F2 plane and vowel /**u**/.

The boundaries determined has been tested simultaneously for all three planes against the each 10ms window sequence at the output of **fmanal** application and we have got preliminary results (as it would be presented at Results). The percentage of wrong recognized vowels was high and we have decided to put additional boundaries as it is presented at Figure 8. If we consider the overlapping of vowels /**a**/ and /**e**/ in F1-F2 plane (Fig. 8a) we can notice that we can put the line between the points of intersection of areas that vowels occupy. We have put demarcation lines to vowels /**a**/, /**e**/ and /**i**/ in F1-F2 plane (Fig. 8b) and to vowels /**a**/, /**o**/ and /**u**/ in F2-F3 plane. In this way the overlapping is minimized.

In addition to these criteria for recognition of vowels, it is necessary to remove the consonants. This has been done by considering the amplitude of formants and fundamental frequency.

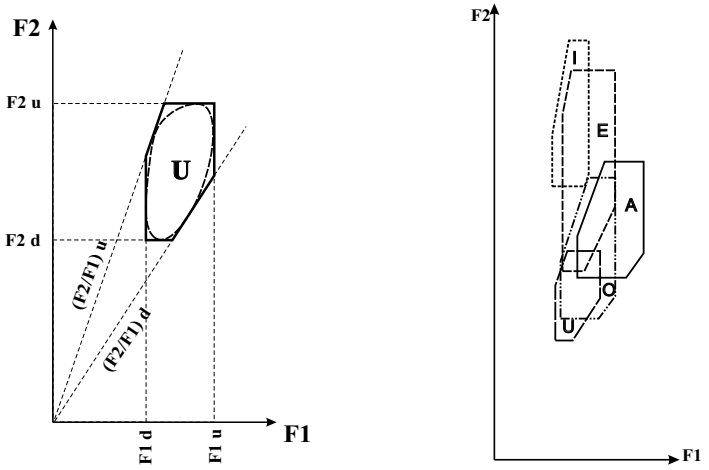


Fig. 7. Boundaries of vowels in F1-F2 plane. (a) Determination of boundaries in F1-F2 plane. (b) Overlapping of boundaries for all vowels in F1-F2 plane.

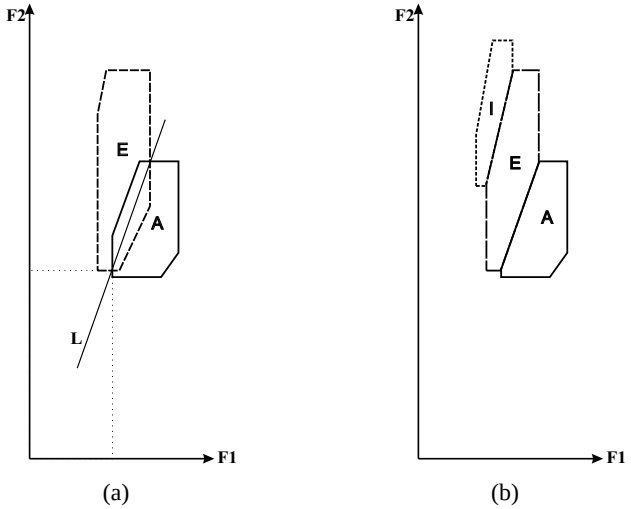


Fig. 8. Additional boundaries of vowels in F1-F2 plane. (a) Determination of additional boundaries in F1-F2 plane. (b) Non-overlapping boundaries for vowels in F1-F2 plane

We have rejected all the time sequences where the fundamental frequency at the output of application is 0 (in order to reject the voiceless consonants /**p**/, /**t**/, /**k**/, /**ć**/, /**s**/, /**š**/, /**f**/, /**h**/ and /**c**/).

In order to eliminate as many sequences of voiced consonants as possible, am-

plitudes of formants are multiplied with weighting factors:

$$A_v = A_1 \times 2^2 + A_2 \times 2^1 + A_3 \times 2^0 \tag{6}$$

and we have noticed that such a sum of formant amplitudes is significantly lower for consonants than for vowels. We have adopted the lower bound on Av_{min} that occupy about 99% of vowel sequences. We have considered as consonants all sequences with $Av < Av_{min}$. The distribution of Av calculated for all sequences is presented at Figure 9.

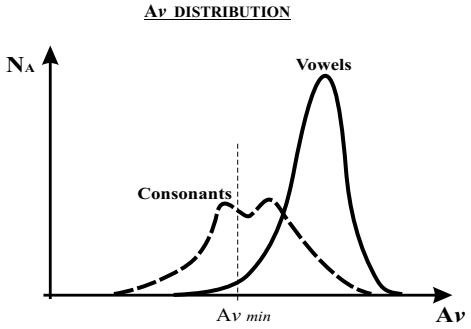


Fig. 9. Distribution of function Av for consonants and vowels.

4.1 Selection and pre-processing of speech signal

Speech signal has been taken from the database of speech signals that was created 1990th at the Faculty of Electrical Engineering in Belgrade. This database contains the speech signals of 60 speakers (30 male and 30 female). The speakers were students who had no obvious speech defects. The paragraph of text was read in normal rhythm of speech.

Spoken material was directly recorded in the Electro-Acoustic laboratory of the Faculty. The high-quality microphone was used and a hardware filter has limited the signal to the frequency range from 200 Hz to 8000 Hz. The signal was imported in WAV format by SFS program, and then re-sampled to 10000Hz. Re-sampling was done to highlight the peaks at frequencies below 5 kHz, which are most relevant for the analysis. Annotation (markers) were determined by audio-visual method. We have chosen the speech signal spoken by three male and three female speakers to perform our formant analysis.

As already mentioned, the application that has been used is **fmanal**, with a fixed window width of 20 ms and overlapping of 10 ms. As a result of every 10 ms processing, coefficients has been obtained that represent: the frequency, bandwidth and amplitude of the first three formants.

5 Results and Discussion

We have tested the recognition of vowels at sequences that we have got at the output of **fmanal** application. We have divided our results in two parts. In the first part we present the results of recognition of vowels on sequences already marked as vowel sequences (sequences where vowels are present in annotations). The results of recognition of vowels applied to all sequences are presented at second part.

5.1 Recognition of vowels applied to sequences marked as vowels

The recognition matrix of vowels by using formant frequencies when areas of vowels are overlapping in F1-F2-F3 space is shown at Table 3.

Table 3. Recognition matrix when areas of vowels are overlapping in space F1-F2-F3.

vowels	A	E	I	O	U
A	**81.3**	**14.6**	0	**16.6**	4.2
E	**15.2**	**83.3**	**69**	6.1	0.4
I	0.3	**36**	**79.8**	0.5	0
O	**18.1**	6	0.5	**80.9**	**52**
U	0.6	1.3	0	**36**	**80.1**

It can be seen from Table 3 that the correct recognition of vowels (in diagonal cells) is about 80%. But, there are also wrong recognitions that are significant (E-I 69%, O-U 52%, I-O and U-O 36%). The wrong recognitions are expected because of large overlapping of formant frequencies for some vowels and it can be seen at Figure 7b. It was the reason why we added demarcation lines in F1-F2 and F2-F3 planes for some vowels which space overlapping was large.

After introduction of additional delineation of vowel areas in formant space, we have obtained the recognition results as shown at Table 4.

Table 4. Recognition matrix when areas of vowels are demarcated in space F1-F2-F3.

vowels	A	E	I	O	U
A	**68.1**	**3.7**		**6.14**	
E	**10.3**	**68.33**	**25.5**		
I		**17.33**	**64.88**		
O	**13.59**			**54.1**	**28**
U				**17.1**	**53**

It can be seen in the Table 4 that the performance of recognition system is de-

creased by using demarcation lines in F1-F2 and F2-F3 planes. The percentage of correct recognition is decreased on average of 61.67%. But the percentage of wrongly recognized vowels is decreased too. We have decreased the top percentages of 69% (E-I), 52% (O-U), 36% (I-E and U-O) to 25.5%, 28%, 17.33% and 17.1% respectively. Once more to say, we have not put demarcation lines everywhere (for all vowels at every plane), but only where is the sense. For example, we couldn't put delineation in F1-F3 plane because the overlapping was too high (Fig. 5c).

5.2 Recognition of vowels applied to all sequences

When the recognition of vowels has been tested on all sequences, we have adopted the rule that the recognition is successful if two consecutive sequences points to the same vowel. In this way, we have got the results and errors as shown at tables 5 and 6.

Table 5. Recognition of vowels in all sequences.

vowel	number of vowels in text	recognized	not recognized	percentage
A	41	37	4	90.24
E	30	25	5	83.33
I	34	29	5	85.29
O	34	27	7	79.41
U	9	7	2	77.78

Table 6. Error rates in recognition of vowels in all sequences

vowel	total number of phonems in text	errors	percentage
A	341	18	5.28
E	341	30	8.80
I	341	10	2.93
O	341	22	6.45
U	341	13	3.81

Using the adopted rule that the recognition is considered as successful if two consecutive sequences are recognized as sequences of same vowel, the percentage of correct recognized vowels has become better and it can be seen at confusion matrix at Table 7.

Table 7. Confusion matrix in recognition of vowels in all sequences.

vowels	A	E	I	O	U
A	**90.24**	**16.6**	0	**8.82**	11.1
E	**9.76**	**83.33**	**23.53**	6.06	11.1
I	0	**16.6**	**85.29**	0	0
O	**17.07**	0	0	**79.41**	**22.22**
U	0	0	0	**26.47**	**77.77**

6 Conclusion

In this paper we have presented a simple method for recognizing the five vowels of the Serbian language in continuous speech. The method we have used is based on recognition of frequencies of first three formants that are present in vowels. By using of LPC method for determining the frequencies and amplitudes of formants in speech, we have set the frequency ranges of formants F1, F2 and F3 for all vowels and defined the areas that vowels occupy in F1-F2-F3 space. The areas of vowels in F1-F2-F3 space overlaps, and it has been the reason that we have got a large percentage of wrongly recognized vowels. By introducing the demarcation lines in F1-F2 and F2-F3 planes for some vowels, in order to make non-overlapping areas for recognition, we have obtained less wrongly recognized vowels, but also the correct recognition rate has been reduced.

The best results we have realized are achieved by using the rule to consider vowel recognized if it is recognized in at least two consecutive time windows. When recognition is performed only to vowel speech samples, the average correct recognition rate we have obtained was 83.2% (90.24% the best and 77.77% the worst), and the largest wrongly recognized vowel percentage was 26.47%. But, when the recognition of vowels has been performed on whole speech signal, the average error rate was 5.45%.

These results leads us to conclude that the described algorithm can be implemented to systems for Automatic Speech Recognition, especially for recognition of vowels in continuous speech of Serbian language. This algorithm can be easily applied to other languages too.

References

[1] S. T. Jovičić, *Govorna komunikacija, fiziologija, psihoakustika i percepcija*. Izdavačko preduzeće Nauka, 1999.

[2] L. Rabiner and B. Juang, *Fundamentals of speech recognition*. Prentice Hall, 1993.

[3] J. Holmes, "The JSRU channel vocoder," *Proceedings IEE*, vol. 127, 1980.

[4] S. Furui, *Digital Speech Processing, Synthesis and Recognition*. New York and Basel: MARCEL DEKKER, INC., 1989.

[5] H. Iqbal, M. Awais, S. Masud, and S. Shamail, "On vowels segmentation and identification using formant transitions in continuous recitation of quranic arabic," in *New Challenges in Applied Intelligence Technologies*, ser. Studies in Computational Intelligence. Springer Berlin / Heidelberg, 2008, vol. 134, pp. 155–162.

[6] Y. A. Alotaibi and A. Hussain, "Comparative analysis of arabic vowels using formants and an automatic speech recognition system," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 3, 2010.

[7] S. A. M. Yusof, P. M. Raj, and S. Yaacob, "Speech recognition application based on malaysian spoken vowels using autoregressive model of the vocal tract," in *Proceedings of the International Conference on Electrical Engineering and Informatics*. Bandung, Indonesia: Institut Teknologi Bandung, June 2007.

[8] G. N. Kodandaramaiah, M. N. Giriprasad, and M. M. Rao, "Independent speaker recognition for native english vowels," *International Journal of Electronic Engineering Research*, vol. 2, 2010.

[9] D. A. Kocharov, "Automatic vowel recognition in fluent speech," in *Proceedings of the 9th Conference of Speech and Computer*, St. Petersburg, Russia, Sept. 2004.

[10] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, 2005.