Vowel recognition using formant analysis and neural networks

Emilija Kisić, Goran Dikić and Vera Petrović

***Abstract*—In this paper, a system for vowel recognition using formant analysis and neural network is described. Complete procedure for vowel recognition which consists of historical dataset forming, dataset preprocessing, power spectral density estimation, formant extraction and neural network training and testing is given. Finally, gain results are discussed and it is shown that with first three formant frequencies and with appropriate neural network architecture vowels can be classified and recognized with big accuracy.**

***Index Terms*—formant frequencies; vowel recognition; neural network.**

I. Introduction

Automatic speech recognition (ASR) is scientific field which attracts scientist and researches for more than 60 years. Full development of this scientific field has happened with the transition from analog to digital systems. Recently, with global technological development, ASR has gained application in large number of applications that can be found in everyday life [1].

In this paper automatic vowel recognition using formant analysis and neural network is described. It is known from acoustic theory of speech production, that every uttered vowel has three main resonant frequencies which are called formant frequencies or just formants [2, 3]. In this paper is described complete procedure for automatic vowel recognition. First step was forming of dataset which consists of recorded vowels uttered by male speakers. After that, dataset was processed for noise cleaning and for extraction of useful part of every recorded vowel. After dataset preprocessing, power spectral density estimation of signals is performed, so formants could be extracted [4]. After power spectral estimation, first three formant frequencies were extracted from each vowel using adaptive algorithm. Extracted formants were used for neural network training [5, 6]. After training, neural network was tested on new vowels that were not in historical dataset. Very good results were gained during training and during neural network testing. System for automatic vowel recognition using formant analysis and neural network gave good results and did recognition with big accuracy. It is shown that first three formant frequencies can make good classification between vowels and with good neural network design system can make vowel recognition with big accuracy. On Fig.1 algorithm for vowel recognition is shown.

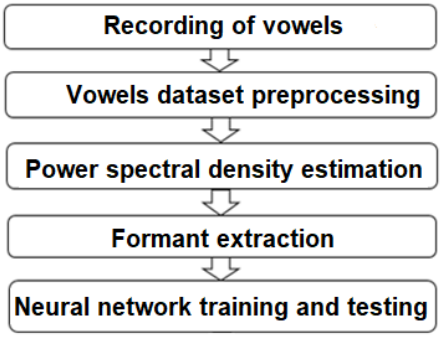

Fig.1. Algorithm for vowel recognition.

First four steps of the algorithm were performed in programming language *Matlab*, while neural network training and testing was performed in programming language *Python.*

II. Acoustic Theory of Speech Production

In acoustic theory term sound refers to vibration. Vibrations are the cause of sound waves production, which propagate due to particle oscillating in the medium through they travel. Because of that, basic principles of physics must describe production and propagation of sound in vocal tract. For vocal tract modeling, detailed acoustic theory must take into account next factors: time variability of the shape of vocal tract, losses due to thermal conductivity and viscous friction on the walls of the vocal tract, softness of the walls of the vocal tract, radiation of sound on the lips, acoustic relation between oral and nasal cavities and sound source in the vocal tract [3]. In this paper simplified mathematical model is taken which neglects the above factors. The simplest model which can describe the process of speech production is shown on Fig.2. Vocal tract is modeled as a tube of uneven, time-varying cross-section.

Frequency characteristic of vocal tract is defined as ratio between the complex amplitude of the volumetric air flow at the end and the beginning of the tube. In the field of ASR resonant frequencies of vocal tract tube are called formant frequencies or just formants.

Emilija Kisić is with the School of Electrical and Computer Engineering of Applied Studies, 283 Vojvode Stepe, 11000 Belgrade, Serbia (e-mail: emilija.kisic@ viser.edu.rs).

Goran Dikić is with the School of Electrical and Computer Engineering of Applied Studies, 283 Vojvode Stepe, 11000 Belgrade, Serbia (e-mail: gdikic@ viser.edu.rs).

Vera Petrović is with the School of Electrical and Computer Engineering of Applied Studies, 283 Vojvode Stepe, 11000 Belgrade, Serbia (e-mail: verap@ viser.edu.rs).

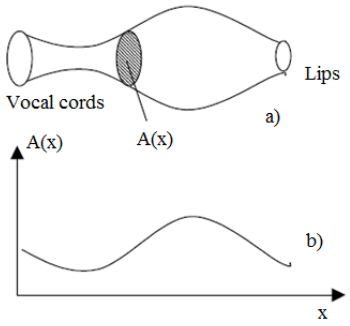

Fig.2. a) Vocal tract model b) Transfer function of vocal tract cross-section, $A(x,t)$.

Formant frequencies depend on the shape and dimensions of the vocal tract. Therefore, the spectral properties of the speech signal change over time with the change in the shape of the vocal tract. The bandwidth of the first formant (first formant frequency) is mainly determined by the vibration of the walls. The bandwidths of the second and third formant frequencies are determined with the combination of the effects of wall vibrations and radiation on the lips. At high frequencies, the influence of radiation on the lips is dominant, which at these frequencies overcomes the influence of wall vibrations, friction and thermal conductivity. Information about the content of the speech signal is hidden in the first two formants, while the third and fourth formant refers to the color of the voice [3].

III. Recording of Vowels

The first step in making a system for vowel recognition was vowel recording. Vowels are voices that arise with quasi-periodic excitation, where the function of cross-section of the vocal tract is stationary. The way in which the cross section of the vocal tract changes determines the resonant frequencies of the tract (formants) and thus the sound is produced. Each vowel can be characterized by the function of cross-section of the vocal tract used in its production. It is obvious that this is very imprecise characterization due to the natural differences that exist between the vocal tracts of different speakers. The second representation is trough the resonant frequencies of the vocal tract. Also, in this case, there are numerous variations that are expected for the same vowel that is uttered by a large number of different speakers. The period of the basic frequency of oscillation of the vocal cords, i.e. the period of the glottal wave is called the pitch period. It ranges from $(7 \div 8)ms$ for men, about $(4 \div 5)ms$ for women and about $(2.5 \div 3.5)ms$ for children. In other words, the fundamental frequency of vocal cord is about $(100 \div 120)Hz$ for men, about $(200 \div 250)Hz$ for woman and about $(300 \div 350)Hz$ for children. For this reason, the formant frequencies for female speakers and children are shifted relative to the formant frequencies for male speakers. In Table I are shown first three average formant frequencies for vowels uttered by male (native English) speakers [3].

TABLE I
Formant Frequencies for Male Speakers

First three formant frequencies for vowels pronounced by male speakers			
Vowel	F_1	F_2	F_3
A	730	1090	2440
E	530	1840	2480
I	390	1990	2550
O	570	840	2410
U	300	870	2240

Because of different values of formant frequencies for male and female speakers and children it would be a very difficult task to make a unique algorithm for vowel recognition for vowels uttered by men, women and children. For this reason we decided to make an algorithm for vowel recognition uttered only by male speakers. Original base of vowels consisted of 500 uttered vowels. Ten male speakers uttered each vowel for 10 times. Recording of each vowel lasted 2 seconds with sample frequency of 8 kHz. The sample frequency was chosen in order to satisfy Shannon's sampling theorem [7].

IV. Dataset Preprocessing

After recording of vowels, next step was dataset preprocessing. All vowels were filtrated for removing of noise and high frequencies that are not of interest for first three formant frequencies seeking. On Fig.3 speech signal which represents uttered vowel "a" after filtering is shown.

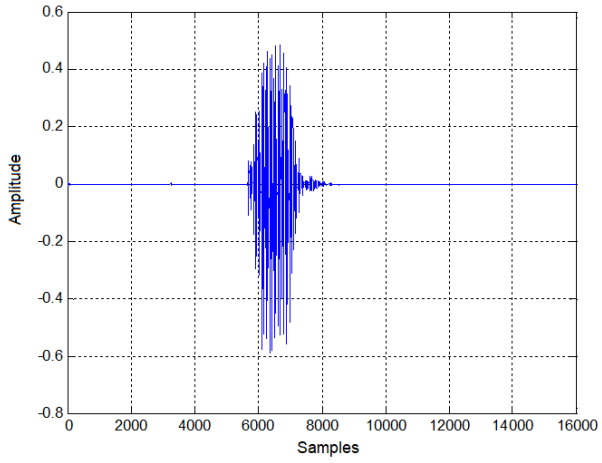

Fig.3. Uttered vowel "a" after filtering.

After filtering, it was necessary to extract only useful part from recorded signals. As we can see from Fig.3 beginning and the end of recorded signal is useless for further analysis, because it is part of a signal which represents silence. Good solution for extraction of useful part of the speech signal is calculating the square of amplitude of signal and extracting the part which is above some threshold [3]. The original

signal with squared amplitude can be defined with equation:

$$X = x(n)^2, \tag{1}$$

where n is number of samples. Part of each recorded signal that was above the threshold which is 25% of the maximum of signal with squared amplitude was extracted. Threshold was chosen empirically. In this way, only useful part of the signals was extracted and useless part was removed (part at the beginning and the end of the signal which represents silence). On Fig.4 speech signal which represents uttered vowel "a" with squared amplitude and threshold are shown.

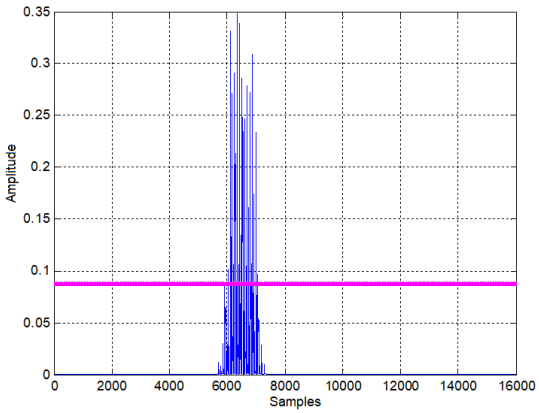

Fig.4. Speech signal which represents uttered vowel "a" with squared amplitude and threshold.

The useful part of each recorded vowel was divided into "packets" of 800 samples and thus, after preprocessing a dataset consisting of 1080 signals is obtained. Dividing of useful part of the signals is performed for increasing of dataset of uttered vowels.

V. Power Spectral Density Estimation of Speech Signals

In order to extract formant frequencies, it was necessary to perform power spectral density estimation of signals. Power spectral density (which is noted as $P_{xx}(f)$) of complex, wide stationary random process $x[n]$ is defined as:

$$P_{xx}(f) = \sum_{k=-\infty}^{\infty} r_{xx}[k]\exp(-j2\pi fk), \quad -\frac{1}{2} \le f \le \frac{1}{2}. \tag{2}$$

With $r_{xx}[k]$ is noted autocorrelation function of $x[n]$ defined as:

$$r_{xx}[k] = \varepsilon(x^*[n]x[n-k]), \tag{3}$$

where ε represents mathematical expectation operator.

The power spectral density function actually represents the distribution of power over the frequency of a random process. Since the power spectral density is a function of an infinite number of values of the autocorrelation function, the task of estimating the power spectral density based on a finite set of data is almost impossible. There are various models for power spectral density that can be assumed in order to minimize the problem of spectral estimation. The choice of a model may depend on which model best finds the required spectral characteristics. An example of this is the search for formant frequencies in speech signal. Spectral estimation via Welch [8] gave the best results comparing with other two methods-Periodogram and Blackman-Tukey [4].

Basically, Welch's method is based on time averaged periodogram which is defined as:

$$\hat{P}_{PER}(f) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x[n]\exp(-j2\pi fn)\right|^2, \tag{4}$$

where N is number of samples. According to Welch's method, the number of samples N in which we perform the estimation should be divided into L intervals of M samples and each of these intervals should be multiplied by a window. For each interval multiplied by a window, a periodogram should be calculated and at the end averaging of periodograms should be performed. Multiplying with a window solves the problem of "spectrum leakage", i.e. prevents the occurrence of signals at lower levels to be masked by the side lobes of signals at higher levels, if the signals are close in frequency. Multiplying with the window reduces the signal level on the side lobes, at the cost of increasing the width of the main lobe. Welch's method solves the problem of making compromises between spectral resolution, variance and bias by allowing data intervals to overlap. As the N increases, variance decreases, the estimation is not shifted, and since the intervals overlap, we did not lose much when it comes to resolution. In this paper, Hamming window [7] was chosen with length of 128 samples, and the overlap between data intervals was 50%. On Fig.5 is presented power spectral density estimation of uttered vowel "a" via Welch.

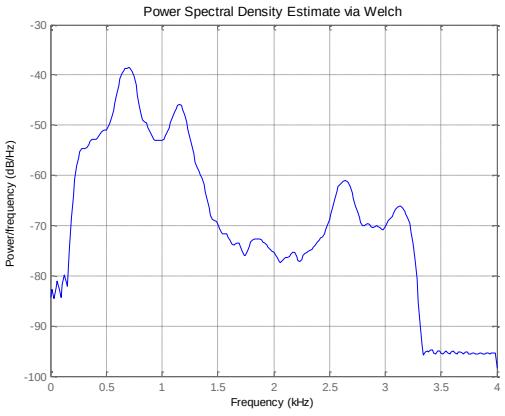

Fig.5. Power spectral density estimate via Welch for uttered vowel "a".

From Fig.5 can be noticed that formant frequencies are very clear represented as peaks in the signal, so they can be easy extracted. First formant is always global maximum,

while second and third formants are local maxima.

VI. Formant Extraction

After dataset preprocessing all recorded vowels were filtered, useful part of the signals was extracted and divided in smaller "packets" and for dataset consisting of 1080 preprocessed signals power spectral density estimation was performed. In this way dataset was prepared for formant extraction.

An adaptive algorithm was developed, based on Table I, depending on the range in which the formant frequencies are expected to appear. The first formant was found as the global maximum, and the remaining two formants were found based on where we expected them to appear, depending on the vowel. Only for the vowel "a" the range for the first formant does not overlap with the ranges for the first formant of other vowels, so it was easiest for the vowel "a" to create a separable class. The overlap of the first formant frequencies occurs with the vowels "e" and "i", but they can be classified quite successfully based on the position of the second and third formant frequencies. Also, the vowels "o" and "u" overlap by the second formant, but they can be classified based on the position of the first and third formant frequencies. Another difficulty that occurred during creating an adaptive algorithm was that formant frequencies do not always appear in expected ranges. Formant frequencies can be shifted higher or lower than expected, depending on the uttering that can vary for different speakers. In Table II average values of formant frequencies extracted from the recorded vowels are shown. There are some differences between Table I and Table II in average values of formant frequencies. Table I is formed according to male speakers which are English native speakers, and Table II is formed according to male speakers from Serbia. Another cause of these differences is uttering of vowels. Speakers which utter the vowels are from different age groups, some of them may be smokers, speakers are in different mood during recording, etc. All this affect on formant frequencies positions. Even a same speaker can utter the same vowel differently. This is the reason why ASR is very difficult task.

TABLE II
AVERAGE VALUES OF FORMANT FREQUENCIES

Average values of formant frequencies extracted from recorded vowels pronounced by male speakers			
Vowel	F_1	F_2	F_3
A	744	1177	2591
E	442	1889	2583
I	335	1790	2681
O	468	840	2518
U	354	865	2442

After formant extraction, five classes were obtained, one class for each vowel, but their overlapping could not be avoided. Classification in five separate classes based on first three formant frequencies is difficult because of formant frequencies overlap between some vowels, but the adaptive algorithm that was applied to extract the formants gives quite good results which are shown on Fig.6.

After formant extraction, for each vowel about 200 formant frequencies (first three formant frequencies) are obtained. This is now prepared historical dataset which will be used for neural network training and testing.

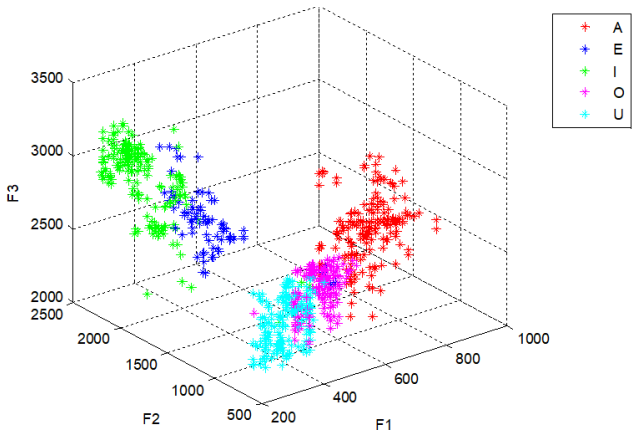

Fig.6. Distribution of the vowels in F_1 F_2 F_3 space.

VII. Neural Network for Vowel Recognition and Gain Results

After formant extraction final step was neural network design. For vowel recognition problem multilayer feed-forward neural network was chosen [9]. Supervised learning technique is performed because neural network has target values for input data. Supervised learning implies that neural network has output values for all input values. On these input/output pairs neural network is trained and tested [10].

Neural network has three inputs, one for each formant frequency. Neural network has five outputs, one for each class that represents one vowel. We decided to split our historical dataset that has 1080 formant frequencies (about 200 instances for first three formant frequencies for each vowel), so 90% of data was used for training and 10% was used for testing. For better algorithm performances k-fold cross validation was performed [5]. Entire dataset was split on 10 folds. For model metrics we used accuracy which was counted on testing dataset in each iteration during k-fold cross validation. Finally, average accuracy of the model was counted for all ten iterations.

Number of hidden layers and nodes was chosen by trial and error method. First, small number of hidden layers and nodes was chosen. This number was increased in order to increase the accuracy of the model, but we took care about overfitting and stopped with increasing when accuracy started to decrease [9]. We also tried network training with different optimization algorithms, activation functions and values of learning rate. Number of epochs was also changed.

The best result was achieved with neural network architecture that consists of six hidden layers with 100 nodes.

Optimization algorithm that was used is Adam [11] with sparse categorical cross entropy loss function, with learning rate one and with RELU activation function [9]. Number of epochs was 2000. With this neural network architecture average accuracy of the model on testing dataset was 96.3%.

When optimal neural network architecture was chosen, neural network was trained on entire historical dataset. Finally, neural network testing was performed on uttered vowels that did not were in original base of vowels. Three new male speakers uttered each vowel five times. For new vowels confusion matrix is shown in Table III.

TABLE III
CONFUSION MATRIX FOR NEW VOWELS

	A	E	I	O	U
A	15	0	0	0	0
E	0	12	3	0	0
I	0	0	15	0	0
O	0	0	0	15	0
U	0	0	0	0	15

New vowels were recognized with accuracy of 96%. We can see that all vowels except vowel “e” were recognized with 100% accuracy. Vowel “e” was three times recognized as vowel “i”. This error is not surprising because classes of vowels for neural network training were not separable, and there was overlapping between classes “e” and “i”. These results could be different for some other speakers and that is the reason why vowel recognition is very complex task.

VIII. CONCLUSION

Considering that historical dataset for neural network training was not big, gain results are very satisfying. It is shown that with first three formant frequencies and with adequate choice of neural network, vowels can be recognized with big accuracy.

First problem during vowel recognition system making was different uttering of vowels. Since this is a system that is independent of the speaker, the way in which vowels are uttered is very important for their recognition. For this reason, it is very difficult to create a single algorithm that classifies vowels, regardless of which speaker uttered the vowel. Second problem was overlapping between classes that represent vowels. This overlapping was unavoidable and it affected on model accuracy.

Dataset with formant frequencies for neural network training was not very big. Increasing of dataset with recorded vowels for neural network training would probably increase model accuracy. Also, more uttered vowels for neural network testing would give more reliable results about model accuracy. Despite the above limitations and problems encountered in designing a vowel recognition system, the designed system gives very good results.

Due to their properties, neural networks are nowadays very attractive for scientists and researchers when it comes to speech recognition [12]. In some future work other types of neural network can be applied for solving vowels recognition problem and comparative analysis can be made with gain results.

REFERENCES

[1] J. Li, L. Deng, R. Haeb-Umbach, Y. Gong, *Robust Automatic Speech recognition-A Bridge to Practical Application*, Oxford, UK: Academic Press, 2016.

[2] L. R. Rabiner, R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632, 1978.

[3] L. Rabiner, B.-H. Juang, B. Yegnarayana,, *Fundamentals of Speech Recognition*, Pearson, India, 2010.

[4] S. Kay, *Modern Spectral Estimation: Theory & Application*, Prentice-Hall, New Jersey, 1988.

[5] C. T. Lin, C. S. G. Lee, *Neural Fuzzy Systems: A Neuro-Fuzzy Synergism to Intelligent Systems*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1996.

[6] J. Tebelskis, “Speech recognition using neural networks”, Ph.D. dissertation, Carnegie Mellon University Pittsburgh, 1995.

[7] S. K. Mitra, *Digital Signal Processing*, A *Computer Based Approach*, Wcb/McGraw-Hill, 4th ed., USA, 2011.

[8] P. D. Welch, “The use of fast Fourier transforms for the estimation of power spectra: A method based on time averaging over short modified periodograms,” *IEEE Transactions on Audio and Electroacoustics*, vol. 15, pp. 70-73, 1967.

[9] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning (Adaptive Computation and Machine Learning Series)*, MIT, USA, 2016.

[10] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, Boston, 1980.

[11] D. P. Kingma, J.L. Ba, “Adam: A method for stochastic optimization,” *arXiv*: 1412.6980v9, 2014.

[12] A. Graves, A. Mohamed, A. G. E. Hinton, “Speech recognition with deep recurrent neural networks” *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada 6645-6649, 2013.