50 ТУ ГАБРОВО ГОДИНИ Journal of the Technical University of Gabrovo, Vol. 50’2014 (44-49)

ИЗВЕСТИЯ
на Технически университет
Габрово

SIGNAL PROCESSING TECHNIQUES FOR SPEECH RECOGNITION

The signal processing group

R. T. Ilarionov [*]
Department of Computer Systems and Technology,
Technical University of Gabrovo, Bulgaria

T. D. Todorov [§]
Department of Mathematics,
Technical University of Gabrovo, Bulgaria

G. S. Tsanev [**]
Department of Computer Systems and Technology,
Technical University of Gabrovo, Bulgaria

S. Y. Yordanov [††]
Department of Automation, Information and Control Systems,
Technical University of Gabrovo, Bulgaria

Received 03 October 2014; Accepted 15 October 2014

Abstract

Most of the authors analyzing speech use the wavelet method for feature extraction and Dynamic Time Warping for comparing two time series in their classifiers. The usage of DTW generates significant difficulties obtaining pattern as an average of series. We introduce an alternative approach without analysis of the cepstral coefficients. The present paper summarizes new methods for recognition of speech. A new finite element signal processing method is demonstrated. The present speech recognition techniques essentially reduce the total computational work, which is very important in the case of a real time application. Successful real life examples are presented. The tests strongly support the considered theory.

Keywords: the finite element method, single command recognition, Quasi–Hermitian interpolant, the mel filter bank.

INTRODUCTION

Speech recognition techniques are an object of a great interest in the last decades. They have a usage in various area of the real life. Undoubtedly, most applications of speech recognition can be found in telecommunications. In particular we list voice enable services, customer care wizards, call center automated attendance, voice access to universal directories and registers etc. Other applications can be found also in automated identification, commands for working with windows and programs controlling wheelchairs and robotics.

Algorithms for signal processing based on wavelet methods [1, 2, 3], splines [4], least square methods [5] are well known. Many authors use Dynamic Time Warping for comparing two signals [6, 7, 8]. Smoothing algorithms are successfully used by S. Steiniger and S. Meyer [9] and R. M. Fernandez – Alcala, J. Navarro – Moreno, J. C. Ruiz Molina and J. A. Espinosa – Pulido [10].

We introduce a new approach processing each input phoneme by a particular frequency. The present paper summarizes an original finite element signal processing method for voice command recognition. A smoothing algorithm based on a Quasi-Hermitian interpolant is obtained. The finite element method is successfully used in the classifier, where author`s score functional is defined.

Matlab software product has been used for all of the simulations herein. The present feature extraction methods assure almost a hundred percentage of recognition of the tested phonemes. A lot of real life experiments demonstrate the advantages of the present approach.

The paper is organized as follows. The finite element method for signal processing is described in the next section. Further, real life experiments for single voice recognition by finite element analysis are discussed. In the end concluding remarks are presented.

FINITE ELEMENT METHOD FOR SPEECH RECOGNITION

Suppose that we have to control a system by single commands containing no more than four words. A single voice command is presented by a set of samples $l = \{l_i, i = 1,2,\dots,\sigma\}$. The input command consists of a set of words. Finding samples with an amplitude more than 3% determines the beginning of each word. First sample with an amplitude less than 3% fixes the end of the corresponding word. Consider the processing of a single word $\underline{d}$ since all words in a command are processed by the same way. Prepare a pattern $\underline{d}_p$ of the word $\underline{d}$ recorded at frequency 44.1 kHz. Each input signal is processed by a particular frequency.

Calculate the corresponding frequency $(\nu_s)_I$ of the word $\underline{d}$ by an original conversion formula

[*] Тел.: 066801249; e-mail: ilar@tugab.bg

[§] t.todorov@yahoo.com

[**] georgitsanev@gmail.com

[††] sjjordanov@mail.bg

ISSN 1310-6686© 2014 Известия на Технически университет Габрово

$$(\nu_s)_I = \frac{n_I}{n_p}(\nu_s)_p, \qquad (1)$$

where $(\nu_s)_p =$ 44.1kHz, $n_I = \mathrm{Card}(\underline{d}_I)$ and $n_p = \mathrm{Card}(\underline{d}_p)$. Further, we drop the index "I" for the input data but keep the index "p" for the pattern.

Separate commands can be pronounced by different people with different volume. Moreover, the distance from each person to the microphone is not identical. That is why we should normalize the input signal $\underline{d}$ as follows:

$$\underline{\tilde{d}} = \frac{\underline{d}}{\|\underline{d}\|_\infty},$$

where $\|\underline{d}\|_\infty$ is the max norm in $\mathbf{R}^n$.

We need an appropriate window function in order to avoid the spectral leakage. Apply the adjustable window:

$$D_k = \begin{cases} \frac{1-\alpha}{2} - 0.5\cos\frac{2\pi k}{n} + \frac{\alpha}{2}\cos\frac{4\pi k}{n}, & k = \overline{1,n},\ n = \mathrm{Card}(\underline{d}) \\ 0 & otherwise \end{cases}$$

The adjustable window function makes available the varying of the bandwidth of the main lobe and the side lobe amplitude which cannot be done by a fixed window function. Varying α we can obtain many window functions especially when $\alpha = 0$ and $\alpha = 0.16$. In this particular cases we obtain Hamming and Blackman window correspondingly. The window function becomes Hamming window for $\alpha = 0$, and Blackman one for $\alpha = 0.16$. The following windowed signal is obtained considering all vectors as one column matrices

$$\underline{X} = \mathrm{diag}(\underline{\tilde{d}}^T \underline{D}).$$

Fourier transform is defined as follows:

$$Y_k = \sum_{j=1}^{n} X_j\, e^{-\frac{2\pi(k-1)(j-1)i}{n}}, \quad k = \overline{1,n}.$$

The frequency vector is truncated to the Nyquist frequency in order to avoid aliasing.

$$\nu(k) = \frac{k\nu_s}{n}, \quad k = \overline{0,\tfrac{n}{2}},$$

where ν_s is the frequency obtained in (1).

Mel scale is the best one for speech recognition as it is very close to the frequency response of human auditory system. Present the expression converting from Herz to Mel:

$$T(\nu) = 2595\log_{10}\left(1 + \frac{\nu}{700}\right).$$

We need a mel filter bank in order to obtain the mel frequency cepstrum. Denote the lower and the higher boundaries of the filter bank by ν_* and ν^*. Define

$$\Phi_k(\nu) = \begin{cases} 0 & \nu \le \nu_{k-1} \\ \frac{\nu - \nu_{k-1}}{\nu_k - \nu_{k-1}} & \nu_{k-1} \le \nu \le \nu_k \\ \frac{\nu_{k+1} - \nu}{\nu_{k+1} - \nu_k} & \nu_k \le \nu \le \nu_{k+1} \\ 0 & \nu \ge \nu_{k-1} \end{cases}, \quad \nu_k = \nu(k),$$

where the central frequency of the k-th filter band is computed by

$$\nu_k = \frac{n}{\nu_s} T^{-1}\left(T(\nu_*) + k\frac{T(\nu^*) - T(\nu_*)}{M+1}\right), \quad k = \overline{1,M}.$$

The output energy of the power spectrum by the mel filter bank is presented as follows:

$$E_k = \sum_{i=\nu_{k-1}}^{\nu_{k+1}} |Y_i|^2\, \Phi_k(i), \quad k = \overline{1,M}.$$

Analyzing the logarithmic energy vector

$$\underline{U} = \{\ln E_k,\ k = \overline{1,M}\}$$

we do not make analysis of the extracted features in the time domain. All conclusions in our approach are made on the results in frequency domain which means that the computation of the cepstral coefficients is not necessary. This essentially reduces the overall computational work.

Applying finite element theory we should introduce some functional spaces. Let J be a closed interval in $\mathbf{R}$. We denote the real Sobolev space $W_p^k(J)$ for nonnegative integers k and $p = 2$ by $H^k(J)$. The norm and the seminorm in $H^k(J)$ are denoted by $\|\cdot\|_{k,J}$ and $|\cdot|_{k,J}$ correspondingly. Define the set $P_k(J)$ of all polynomials on J of degree not exceeding k. Collect the points $P_k\left(j_k = \frac{k-1}{M-1}, U_k\right)$, $k = \overline{1,M}$.

Suppose that $u(t), t \in J = [0,1]$ is a $C^1(J)$ such that $u(j_k) = U_k$, $k = \overline{1,M}$. We construct a Quasi-Hermitian interpolant $I_h u$, $h = \frac{1}{M-1}$ taking into account that $u(t)$ is an unknown function. Note that the explicit presentation of the function u is not necessary to be known. It is just enough to suppose that u is sufficiently smooth.

Let $\underline{j}$ be a set of nodes of a uniform finite element triangulation $\tau_h = \{J_k = [j_{k-1}, j_k], k = \overline{1,M}\}$ of the interval J. We have reference finite element $\hat{J}$ which corresponds to the target interval J. Define the approximate derivative $D_h u$ on $\underline{j}$ as follows:

$$D_h u_0 = \frac{1}{6h}(-11u_0 + 18u_1 - 9u_2 + 2u_3),$$

$$D_h u_1 = \frac{1}{6h}(-2u_0 - 3u_1 + 6u_2 - u_3),$$

$$\dots\ \dots\ \dots\ \dots\ \dots\ \dots\dots\ \dots\ \dots\dots\ \dots\ \dots$$

$$D_h u_k = \frac{1}{12h}(u_{k-2} - 8u_{k-1} + 8u_{k+1} - u_{k+2}), \qquad k = \overline{2, M-2},$$

$$\dots\ \dots\ \dots\ \dots\ \dots\ \dots\dots\ \dots\ \dots\dots\ \dots\ \dots$$

$$D_h u_{M-1} = \frac{1}{6h}(u_{M-3} - 6u_{M-2} + 3u_{M-1} + 2u_M),$$

$$D_h u_M = \frac{1}{6h}(-2u_{M-3} + 9u_{M-2} - 18u_{M-1} + 11u_M).$$

Hermitian nodal basis functions are introduced by:

$$\hat{\varphi}_1(\hat{t}) = 2\hat{t}^3 - 3\hat{t}^2 + 1, \quad \hat{\psi}_1(\hat{t}) = \hat{t}(1-\hat{t})^2,$$

$$\hat{\varphi}_2(\hat{t}) = 3\hat{t}^2 - 2\hat{t}^3, \quad \hat{\psi}_2(\hat{t}) = (\hat{t}-1)\hat{t}^2$$

We construct a Quasi-Hermitian finite element space as follows:

$$V_h = \left\{ v_h = \sum_{K\in\tau_h}\sum_{i=1}^{2} \left(v_h(t_{Ki}), D_h v_h(t_{Ki})\right)\cdot\left(\varphi_{Ki}, \psi_{Ki}\right) \middle| \varphi_{Ki}|_K, \psi_{Ki}|_K \in P_3(K), \qquad K \in \tau_h \right.$$

Define the Quasi-Hermitian interpolant $I_h : C^1(J) \to V_h$ by

$$I_h v = \sum_{K\in\tau_h}\sum_{i=1}^{2} \left(v(t_{Ki}), D_h v(t_{Ki})\right)\cdot\left(\varphi_{Ki}, \psi_{Ki}\right). \qquad (2)$$

where φ_{Ki}, ψ_{Ki} $i = 1,2$ are the nodal basis functions of the finite element $K \in \tau_h$ associated with the nodes a_{1K}, a_{2K} of K. The basis functions satisfy:

$$\begin{cases} \varphi_1(a_{1K}) = 1 \\ \varphi_1(a_{2K}) = 0 \\ \varphi_1'(a_{1K}) = 0 \\ \varphi_1'(a_{2K}) = 0 \end{cases}, \qquad \begin{cases} \varphi_2(a_{1K}) = 0 \\ \varphi_2(a_{2K}) = 1 \\ \varphi_2'(a_{1K}) = 0 \\ \varphi_2'(a_{2K}) = 0 \end{cases},$$

$$\begin{cases} \psi_1(a_{1K}) = 0 \\ \psi_1(a_{2K}) = 0 \\ \psi_1'(a_{1K}) = 1 \\ \psi_1'(a_{2K}) = 0 \end{cases}, \qquad \begin{cases} \psi_2(a_{1K}) = 0 \\ \psi_2(a_{2K}) = 0 \\ \psi_2'(a_{1K}) = 0 \\ \psi_2'(a_{2K}) = 1 \end{cases}.$$

Present the nodal basis functions φ_{Ki}, ψ_{Ki} by the nodal functions of the reference element $\hat{J}$. The functions φ_{Ki}, ψ_{Ki} and $\hat{\varphi}, \hat{\psi}$ are related by:

$$\varphi_{Ki} = \hat{\varphi}_i \circ F_K^{-1}, \psi_{Ki} = \hat{\psi}_i \circ F_K^{-1} \quad i = 1,2,$$

where F_K is the generating finite element transformation of the finite element $K \in \tau_h$. Then the interpolant (2) can be presented in view of:

$$I_h v = \sum_{K \in \tau_h} \sum_{i=1}^{2} \left(v(t_{Ki}), D_h v(t_{Ki})\right) \cdot \left(\hat{\varphi}_i, \hat{\psi}_i\right).$$

Having in mind that the Euclidean norm is not appropriate for precise comparison of a set of samples we define a score functional with respect to the k-th Sobolev norm

$$Q_h(u) = \frac{E\left(I_h(u - u_p)\right)}{E(I_h u_p)},$$

where E is the energy associated with $\|\cdot\|_{k,J}$. The higher order energy functional Q_h can be successfully used in $H^k(J)$, $k = 0,1$. But in the case $k > 1$ a smoother interpolation operator should be constructed.

If the score functional satisfies $Q_h(u) < \mu$ we have recognition of a certain word. We set experimentally $\mu = 5\%$.

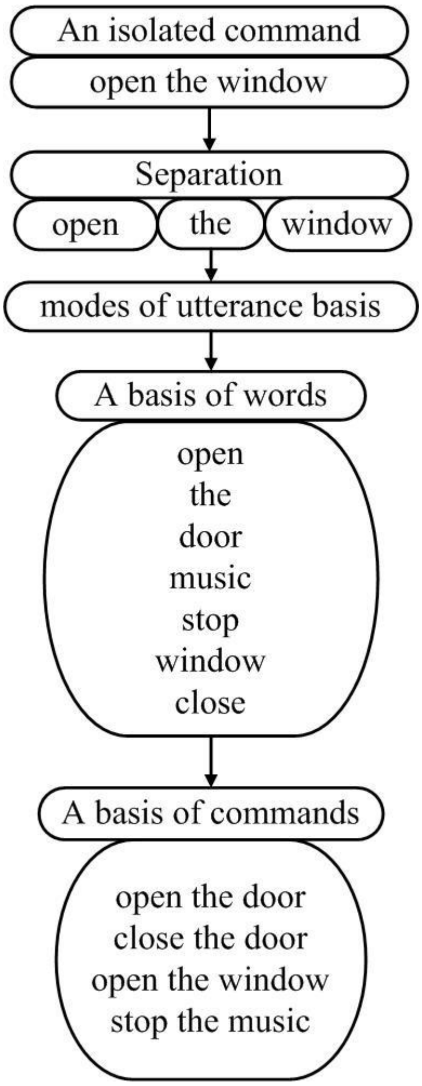


Figure 1. An example for recognizing a single voice command.

We illustrate the algorithm by a simple example, presented in Figure 1. There are various pronunciations of a single word. Consider for example the word “club”. It could be said as [klub], [klʌb] or [kləb]. Therefore, we need a modes of utterance basis. Consider a single voice command "open the window". The command consists of three words. The processing of the command starts searching for the first word in the basis of utterances. Finding the input data in the modes of utterance basis we continue searching in the basis of words. After finding the word "open" it returns to the second word in the considered command. Search again into the basis of utterances. Recognizing the second word of the command we look for commands starting with "open the" in the basis of commands. We establish that there are two commands "open the door" and "open the window" satisfying the criterion. That is why we analyze the next word in the input command. Finally, the third word "window" is found and the correct command should be recognized from the basis of commands.

Summarize our algorithm for recognizing of an isolated word. An input frequency should be calculated with respect to the potential pattern. Then we have to reduce the obtained frequency to Nyquist one and to change the scale from Herz to Mel in order to analyze the signal more efficiently. Then we need to find mel frequency cepstrum defining a mel filter bank. On the other hand we process the word windowing the corresponding set of samples. Then we apply fast Fourier transform in order to make analysis in frequency domain. The output energy of the input data is a sum of the Fourier image multiplied by the mel filter bank output. We smooth the logarithmic energy vector by a Quasi-Hermitian interpolant. Finally we have classification block which is responsible for recognizing of the command pronounced. Figure 2 shows a principle block-scheme of speech recognition process.

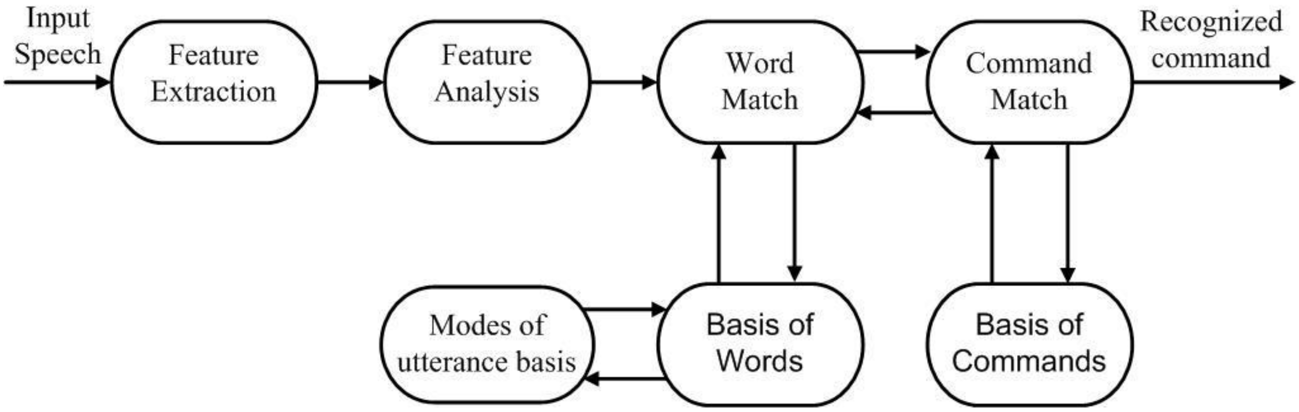
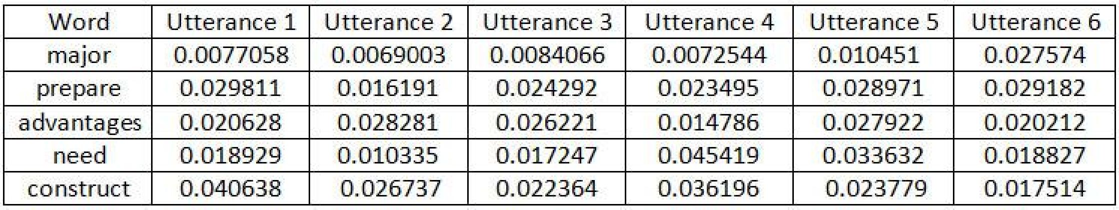


Figure 2. A principle block–scheme of speech recognition process

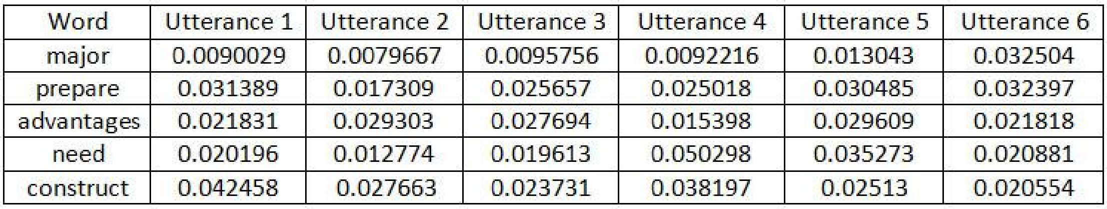
REAL LIFE EXPERIMENTS

A lot of experiments based on the finite element method have been carried out. Various speakers pronounce five different words several times. Analyses with 24, 32, 48 and 64 bands mel filter bank are shown in tables 1-4. As we can see from the tables we have almost a hundred percentage of recognition. In all of the cases shown in the tables can be seen that we obtain results for the score functional much less than 5%.



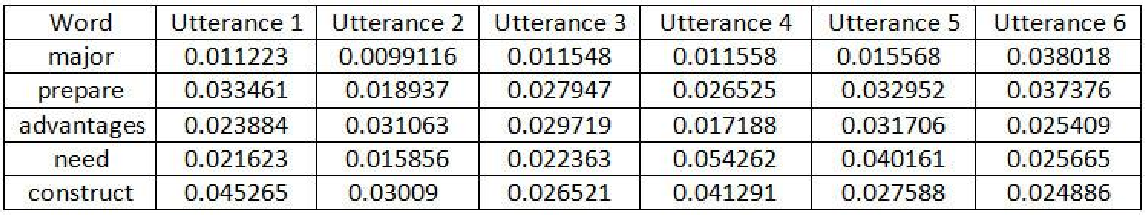
Word	Utterance 1	Utterance 2	Utterance 3	Utterance 4	Utterance 5	Utterance 6
major	0.0077058	0.0069003	0.0084066	0.0072544	0.010451	0.027574
prepare	0.029811	0.016191	0.024292	0.023495	0.028971	0.029182
advantages	0.020628	0.028281	0.026221	0.014786	0.027922	0.020212
need	0.018929	0.010335	0.017247	0.045419	0.033632	0.018827
construct	0.040638	0.026737	0.022364	0.036196	0.023779	0.017514

Table 1 The values of the score functional for *M=24*.



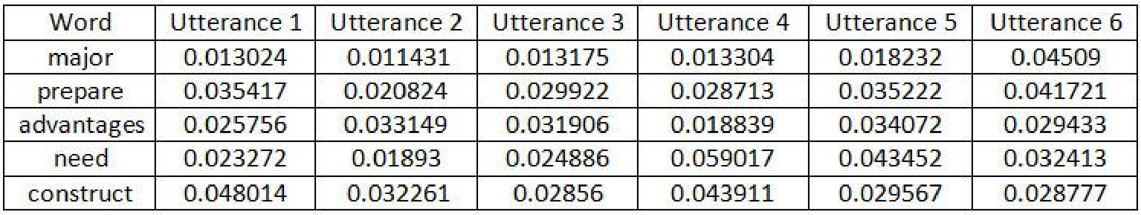
Word	Utterance 1	Utterance 2	Utterance 3	Utterance 4	Utterance 5	Utterance 6
major	0.0090029	0.0079667	0.0095756	0.0092216	0.013043	0.032504
prepare	0.031389	0.017309	0.025657	0.025018	0.030485	0.032397
advantages	0.021831	0.029303	0.027694	0.015398	0.029609	0.021818
need	0.020196	0.012774	0.019613	0.050298	0.035273	0.020881
construct	0.042458	0.027663	0.023731	0.038197	0.02513	0.020554

Table 2 The values of the score functional for *M=32*.



Word	Utterance 1	Utterance 2	Utterance 3	Utterance 4	Utterance 5	Utterance 6
major	0.011223	0.0099116	0.011548	0.011558	0.015568	0.038018
prepare	0.033461	0.018937	0.027947	0.026525	0.032952	0.037376
advantages	0.023884	0.031063	0.029719	0.017188	0.031706	0.025409
need	0.021623	0.015856	0.022363	0.054262	0.040161	0.025665
construct	0.045265	0.03009	0.026521	0.041291	0.027588	0.024886

Table 3 The values of the score functional for *M=48*.



Word	Utterance 1	Utterance 2	Utterance 3	Utterance 4	Utterance 5	Utterance 6
major	0.013024	0.011431	0.013175	0.013304	0.018232	0.04509
prepare	0.035417	0.020824	0.029922	0.028713	0.035222	0.041721
advantages	0.025756	0.033149	0.031906	0.018839	0.034072	0.029433
need	0.023272	0.01893	0.024886	0.059017	0.043452	0.032413
construct	0.048014	0.032261	0.02856	0.043911	0.029567	0.028777

Table 4 The values of the score functional for *M=64*.

The results for the words “need”, “prepare”, “advantages”, “major” and “construct” are presented in figures 3-7. Any particular utterance is colored in blue. Red colored line indicates the corresponding pattern after the training of the system. The process of training consists of obtaining a shape function for each phoneme. For this purpose a linear average operator is used. To obtain an average phoneme by the proposed operator is much easier than to do this by all known Dynamic Time Warping algorithms. This problem becomes much more difficult when Levenshtein distance is applied. Finally in Figure 8 we see the patterns for all five words. They are approximated by the input data.

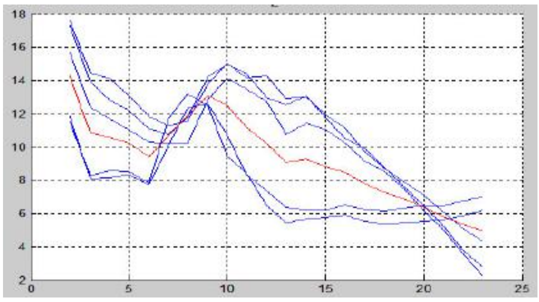


Figure 3. A set of utterances of the word “need”.

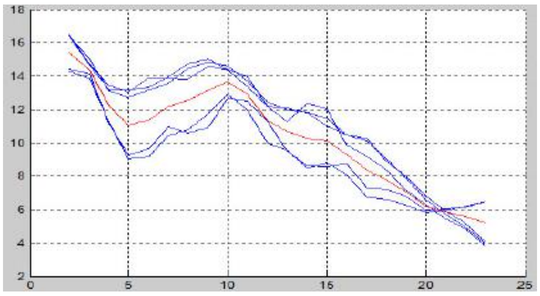


Figure 4. A set of utterances of the word “major”.

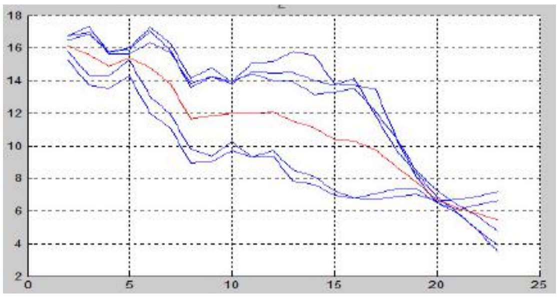


Figure 5. A set of utterances of the word “construct”.

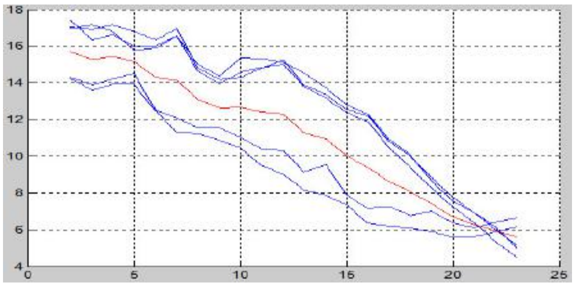


Figure 6. A set of utterances of the word “advantages”.

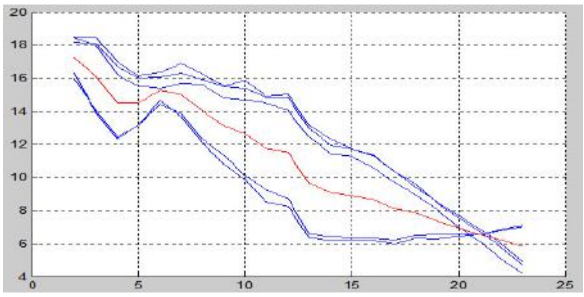


Figure 7. A set of utterances of the word “prepare”.

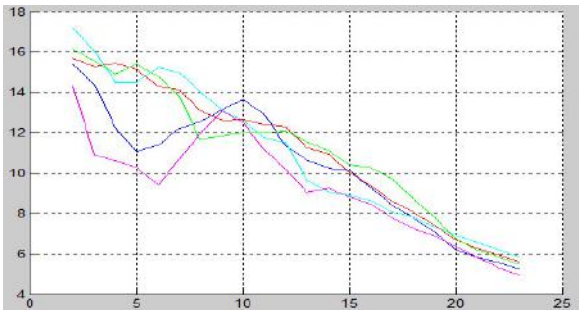


Figure 8. Patterns.

CONCLUSION

A new approach in speech recognition theory is presented. Each input phoneme is processed by the corresponding frequency. A new cubic interpolant of arbitrary C^1 function is obtained. The finite element method is essentially applied in the classifier of the speech recognition process.

The real life experiments indicate that the present method for speech recognition is applicable in any voice controlled system. The proposed approach essentially reduces the overall computational work, which is very important for all real time applications. The implementation of the presented method has been assessed in order to be applied into practice for recognizing speaker independent speech and controlling a wide range of devices including authentication systems.

REFERENCES

[1] W. C. Soares, F. Villarreal, M. A. Q. Duarte, J. V. Filho , "Wavelets in a problem of signal processing", Novi Sad Journal of Mathematics, vol. 41, no. 1, 2011, pp. 11-20.

[2] R. T. Ilarionov, N. A. Shopov, I. S. Simeonov, H. S. Kilifarev, "Ultrasound detection of explosives using wavelets for synthesis of features", vol. 22, no. 8, 2010, pp. 397-407.

[3] M. Unser, P. D. Tafti, "Stochastic Models for Sparse and Piecewise-Smooth Signals", IEEE Transactions on Signal Processing, vol. 59, no. 3, 2011, pp. 989-1006.

[4] M. Unser, A. Aldroubi, M. Eden, "B-Spline signal processing: part II - efficient design and applications", IEEE Transactions on Signal Processing, vol. 41 , no. 2, 1993, pp. 834-848.

[5] N. Ito, W. Qin , "Signal interpolator design using weighted-least-squares method", International Journal of Information and Electronics Engineering, vol. 3, no. 3, 2013, pp. 299-303.

[6] L. Muda, M. Begam, I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing, vol. 2, no. 3, 2010, pp. 2151-9617.

[7] R. Makhijani, R. Gupta, "Isolated word speech recognition system using dynamic time warping", International Journal of Engineering Sciences & Emerging Technologies, vol. 6, no. 3, 2013, pp. 352-367.

[8] A. BALA, "Voice command recognition system based on MFCC and DTW", International Journal of Engineering Science and Technology, vol. 2, no. 12, 2010, pp. 7335-7342.

[9] S. Steiniger and S. Meier, "Snakes: A technique for line smoothing and displacement in map generalisation", in Sixth Workshop on Progress in Automated Map Generalization, Leicester, UK, 2004.

[10] R. M. Fernández-Alcalá, J. Navarro-Moreno, J. C. Ruiz-Molina, J. A. Espinosa-Pulido, "Linear and nonlinear smoothing algorithms for widely factorizable signals", Signal Processing, vol. 93, 2013, pp. 897-903.

[11] S. S. Rajput, S. S. Bhadauria, "Implementation of FIR filter using efficient window function and its application in filtering a speech signal", International Journal of Electrical, Electronics and Mechanical Controls, vol. 1, no. 1, 2012.

[12] A. C. Kelly, C. Gobl, "A comparison of mel-frequency cepstral coefficient (MFCC) calculation techniques", Journal of Computing, vol. 3, no. 10, 2011, pp. 2151-9617.