International Journal of Engineering

Journal Homepage: www.ije.ir

A Clustering-based Approach for Features Extraction in Spectro-temporal Domain using Artificial Neural Network

N. Esfandian*[a], K. Hosseinpour[b]

[a] *Department of Electrical Engineering, Qaemshahr Branch, Islamic Azad University, Qaemshahr, Iran*
[b] *Department of Artificial Intelligence and Robotics, Aryan Institute of Higher Education and Technology, Babol, Iran*

PAPER INFO

Paper history:
Received 07 October 2020
Received in revised form 03 November 2020
Accepted 05 November 2020

Keywords:
Artificial Neural Network
Auditory Model
Clustering
Feature Extraction
Spectro-temporal Features

ABSTRACT

In this paper, a new feature extraction method is presented based on spectro-temporal representation of speech signal for phoneme classification. In the proposed method, an artificial neural network approach is used to cluster spectro-temporal domain. Self-organizing map artificial neural network (SOM) was applied to clustering of features space. Scale, rate and frequency were used as spatial information of each point and the magnitude component was used as similarity attribute in clustering algorithm. Three mechanisms were considered to select attributes in spectro-temporal features space. Spatial information of clusters, the magnitude component of samples in spectro-temporal domain and the average of the amplitude components of each cluster points were considered as secondary features. The proposed features vectors were used for phonemes classification. The results demonstrate that a significant improvement is obtained in classification rate of different sets of phonemes in comparison to previous clustering-based methods. The obtained results of new features indicate the system error is compensated in all vowels and consonants subsets in compare to weighted K-means clustering.

doi*: .5829/ije.2021.34.02b.17*

1. INTRODUCTION

Spectro-temporal representation of the speech signal is considered as one of the important approaches to increase efficiency of speech recognition [1, 2]. One of the limitations of this model is its high dimensional output [3–5]. The output of auditory model is four-dimensional array including the scale, rate, frequency, and time. In recent years, auditory model was used to extract the spectro-temporal features in many applications of speech processing [6–13]. Because of large dimensions of spectro-temporal features space, selection of discriminative features is a crucial task for phoneme classification. Therefore, in recent researches, clustering methods were used to reduce dimension of spectro-temporal features space and extract valuable discriminative information of speech signal. In these methods, output of this model was considered as the primary features vectors and clustered using Gaussian Mixture Model and weighted K-Means [14–17]. Then, the mean vectors and covariance matrices elements of the clusters are considered as secondary features in each speech frame. However, the high computational cost of these methods limited their usability in practical applications. In this article a new method is proposed to extract the discriminative features in spectro-temporal domain. The specific contributions of the manuscript can be described as follows:

In the proposed method, the artificial neural network was used for clustering of spectro-temporal domain according to the capability of artificial neural networks to cluster high-dimensional data [18–21]. The scale, the rate, the frequency and magnitude of each point were assumed as primary features in input vector to cluster using the artificial neural network. The mean vectors, covariance matrices of clusters and the average of the amplitude components of each cluster points were considered as attributes in the secondary feature vectors. The proposed secondary feature vectors were used to classify different categories of phonemes.

*Corresponding Author Institutional Email: *na_esfandian@Qaemiau.ac.ir* (N. Esfandian)

Please cite this article as: N. Esfandian, K. Hosseinpour, A Clustering-based Approach for Features Extraction in Spectro-temporal Domain using Artificial Neural Network, International Journal of Engineering, B: Applications Vol. 34, No. 02, (2021) 452-457

Clustering-based spectro-temporal feature extraction method is briefly discussed in section 2. The proposed secondary feature extraction using the artificial neural network in spectro-temporal domain is presented in section 3. In section 4, the proposed features are experimentally evaluated for phoneme classification of English languages and compared to existing clustering-based approaches. Finally, the paper is concluded in section 5.

2. SECONDARY FEATURES EXTRACTION METHOD IN SPECTRO-TEMPORAL DOMAIN USING CLUSTERING METHOD

The auditory model has two main stages. In the primary stage of this model, an auditory spectrogram was extracted for the input acoustic signal. In the cortical stage, the spectro-temporal features of speech were extracted by applying a set of two dimensional spectro-temporal receptive field (STRF) filters on the spectrogram. The output of cortical stage of auditory model is 4-dimensional vector (scale, frequency, and time). In the clustering-based feature extraction method, the multi-dimensional cortical output was clustered using Gaussian Mixture Model (GMM) and Weighted K-Means (WKM) clustering algorithm. Then, attributes of the clusters such as components of mean and variance vectors were considered as secondary features. Finally, the extracted secondary features were sorted using their estimated weight in the clustering algorithm.

3. SPECTRO-TEMPORAL FEATURES EXTRACTION USING ARTIFICIAL NEURAL NETWORK

In recent researches, the artificial neural network was used to extract the secondary features in various application of speech processing [22–24]. In the proposed feature extraction method, the artificial neural network was used for clustering of spectro-temporal space. The overall architecture of the proposed method illustrated in Figure 1. As it is shown, in the first stage, the auditory spectrogram of the speech signal was computed. Then, in the cortical stage, spectro-temporal modulations were estimated. The auditory spectrogram was obtained using an infinite impulse response (IIR) filter bank with 128 frequency channels between 180 and 7246 Hz at the resolution of 24 channels per octave. In addition, a time constant of 8ms was used for the leaky time integration and filter-bank outputs were sampled every 4 ms to compute the auditory spectrogram. Temporal parameter of the filters (rate) ranging from 2 to 32 Hz and spectral parameter of the filters (scale) ranging from 0.25 to 8 cycle/octave are considered to represent the spectro-temporal modulations of the speech signal. In the next stage, the primary feature vector was extracted using thresholding techniques. The spatial information (scale, rate and frequency) and the magnitude of each point were considered in primary feature vectors. Therefore, primary feature vectors, $v_i = (r_i, s_i, f_i, |A_i|)$, were four-dimensional vector. In this vector, $r_i$denotes the rate, s_i is the scale, f_i is the frequency and $|A_i|$ is magnitude component of each point in spectro-temporal space. These vectors were clustered by the artificial neural network. Eventually, the secondary feature vectors were extracted using the output data of artificial neural network, and used for phoneme classification. In the primary feature extraction algorithm, the amplitude of each sample in spectro-temporal feature space ($|A_i|$) was compared with the maximum amplitude value ($|A_{max}|$) in a speech frame. If the amplitude of each point was higher than an empirically determined threshold value, this point was considered in input vector for clustering. Therefore, valuable discriminative information was only considered in the clustering process. In the proposed secondary feature extraction method, self-organizing map artificial neural network (SOM) was used for clustering of spectro-temporal feature space. SOM is the unsupervised networks. Primary feature vectors, v_i was 4-dimensional vector. Therefore, the input vector of SOM network was a $N \times 4$ matrix which contains N four-dimensional samples. As a result, N four-dimensional input was applied to the SOM network to cluster spectro-temporal domain. In the previous research, three clusters were used for clustering of spectro-temporal feature space in each frame. In this study, assuming three clusters in each frame, three neurons were considered for the artificial neural networks. Block diagram of SOM network is shown in Figure 2. Three-dimensional representation of extracted clusters using SOM network for /g/ phoneme is shown in Figure 3.

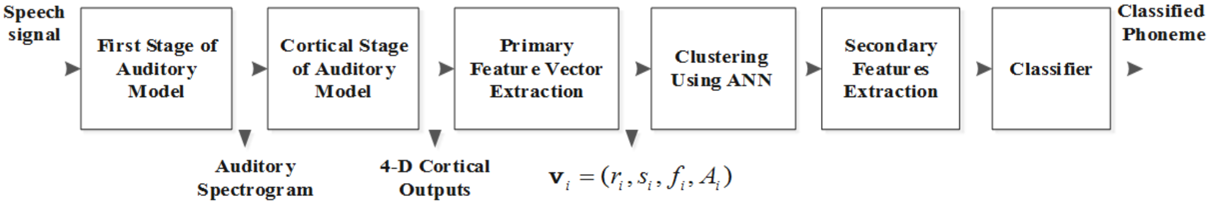
Speech signal
First Stage of Auditory Model
Cortical Stage of Auditory Model
Primary Feature Vector Extraction
Clustering Using ANN
Secondary Features Extraction
Classifier
Classified Phoneme
Auditory Spectrogram
4-D Cortical Outputs
$\mathbf{v}_i = (r_i, s_i, f_i, A_i)$

Figure 1. The overall architecture of proposed feature extraction method

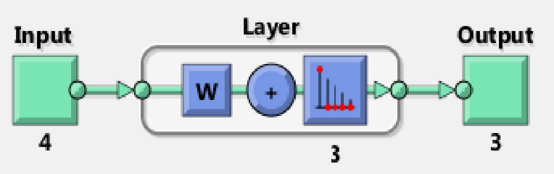

Figure 2. Block diagram of SOM network

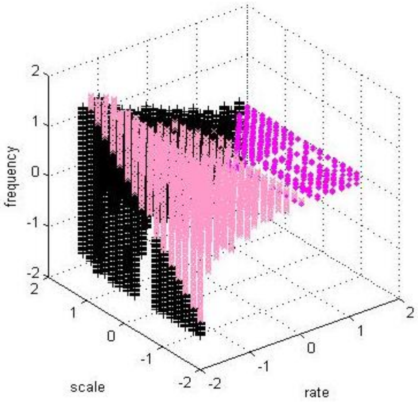

Figure 3. Representation of extracted clusters for /g/ phoneme using SOM network

3. 1. Clustering-based Features Extraction using SOM Network in Spectro-temporal Domain

In proposed secondary feature extraction method, three mechanisms were considered for feature selection in spectro-temporal domain. In the first mechanism, 18 components consist of mean vector $\left(\mu_i = (\mu_{r_i}, \mu_{s_i}, \mu_{f_i})\right)$ and variance vector of clusters $\left(\sigma_i = (\sigma_{r_i}, \sigma_{s_i}, \sigma_{f_i})\right)$ were considered $(V_{SOM,18} = (\mu_i, \sigma_i))$. Spatial information of clusters was considered in this mechanism. The attributes of clusters were sorted with respect to their energy in spectro-temporal space. In the second mechanism, 24 attributes were considered in secondary features vectors $(V_{SOM,24})$. In this feature vector, mean and variance vectors of clusters were also sorted based on energy measure. Each mean or variance vector consists of four components as $\mu_i = (\mu_{r_i}, \mu_{s_i}, \mu_{f_i}, \mu_{A_i})$ and $\sigma_i = (\sigma_{r_i}, \sigma_{s_i}, \sigma_{f_i}, \sigma_{A_i})$. In this mechanism, spatial information and the magnitude component of each point were considered in the secondary features vectors. In the third feature selection mechanism, secondary features vectors consist of 27 attributes $(V_{SOM,27})$. In this mechanism, in addition to 4-dimensional information of clusters, the average of the amplitude components of each cluster points were considered as attribute. The average of the amplitude components of each cluster, $\bar{\mathrm{w}}$, was calculated as follows:

$$\overline{W} = \frac{\sum_{i=1}^{n} w_i}{N} \qquad (1)$$

where, w_i denotes the weight of each point (magnitude component) in spectro-temporal domain. n and N are the number of clusters and the number of points in each cluster.

4. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed method, all types of secondary features vectors (secondary features vectors include 18, 24 and 27 attributes), were used for phonemes classification. The proposed features were evaluated on clean speech from TIMIT database [25]. In this study, SVM classifier was used for phoneme classification [26]. In addition, the optimum values of RBF-SVM parameters (kernel parameter γ and miss-classification cost *C*) were empirically determined using a grid search strategy to optimize the classification rate.

4. 1. Classification Test

The classification error rate of phonemes (/ b /, / d /, / g /) was tabulated for some dialects of TIMIT database using the proposed features in compare to Mel-frequency cepstral coefficients (MFCC) [27] and WKM-based feature vectors (see Table 1).

It is obvious that, the classification results were improved using the proposed features in compare to MFCC and WKM-based features. Also, it is observed that the proposed feature vector consisting of mean and variance vectors of clusters and the average of the amplitude components, $V_{SOM,27}$, gave considerable better results for phoneme classification. Therefore, using the average of the amplitude components of each cluster in the secondary feature vectors improves the phoneme classification rate. Therefore, most of experiments were performed using proposed feature vector including 27 attributes $(V_{SOM,27})$. Confusion matrix for classification of (/b/, /d/, /g/) phonemes using WKM-based features and proposed features in were shown in Tables 2 and 3. It was found that the phoneme /d/ was recognized better than in compare to other phonemes and the greatest error rate was observed for the phoneme /b/ which was recognized incorrectly as /d/.

Simulation was down using MATLAB. Processing time evaluation of proposed features in comparison of MFCC and WKM-based features was shown in Table 4.

TABLE 1. /b/,/d/,/g/ phonemes classification error rate

Dialect	MFCC	WKM	$V_{SOM,18}$	$V_{SOM,24}$	$V_{SOM,27}$
Dialect1	32.1	25.9	24.9	23.3	22.8
Dialect2	35.6	29.2	31.6	29.0	28.1
Dialect3	35.7	28.4	28.1	26.2	26.7
Dialect7	35.1	29.3	27.4	28.4	28.6
Average	35.4	28.2	28.0	26.7	26.6

TABLE 2. Confusion matrix for phoneme classification using WKM-based features

		Recognized		
		b	**d**	**g**
Corrected	**b**	66.9	27.7	5.4
	d	10.1	82.3	7.6
	g	14.9	16.8	68.4

TABLE 3. Confusion matrix for phoneme classification using proposed features

		Recognized		
		b	**d**	**g**
Corrected	**b**	66.5	32.2	1.3
	d	8.9	83.3	7.8
	g	9.1	13.6	77.3

TABLE 4. Processing time evaluation of proposed features in comparison of WKM-based features

	MFCC	WKM	$V_{SOM,18}$	$V_{SOM,24}$	$V_{SOM,27}$
Processing Time (s)	49.1	45.7	32.1	34.5	36.8

As it can be observed, processing time of proposed feature extraction method is less than other features. Phoneme classification results in using the proposed features on different categories of phonemes were compared to multidimensional features (MDF) [28], MFCC and WKM clustering-based features as shown in Table 5. In consonant and vowel subsets, the classification error rate was improved using the proposed secondary features in comparison to MDF, MFCC and WKM-based features. In addition, the relative error improvement in phoneme classification in compare to WKM based features is shown in this table. In vowels, the greatest improvement was obtained for front vowel 50.2 and back vowel 50.4.

TABLE 5. Phonemes classification error rate using MDF, MFCC and WKM-based and proposed features

Phonemes	MDF	MFCC	WKM	Proposed features	Relative improvement (%)
Voiced Plosives (/b/,/d/,/g/)	32.1	38.4	25.9	22.8	11.9
Unvoiced Plosives (/p/,/t/,/k/)	32.3	37.1	31.9	30.3	5.0
Voiced Fricatives (/v/,/dh/,/z/)	16.6	25.9	18.9	16.1	14.8
unvoiced Fricatives (/t/,/s/,/sh/)	12.9	20.3	11.1	8.4	24.3
Nasals (/m/,/n/,/ng/)	49.9	50.3	49.7	42.6	14.2
Front Vowels (/ih/,/ey/,/eh/,/ae/)	43.0	41.5	35.6	17.7	50.2
Back Vowels (/uw/,/uh/,/ow/,/aa/)	29.4	38.1	36.5	18.1	50.4
Diphthongs (/ay/, /aw/, /oy/)	32.7	33.9	28.1	21.6	23.1

5. CONCLUSION

In this paper, clustering-based method was presented for secondary features extraction in spectro-temporal domain. In the proposed method, the spectro-temporal domain was clustered using SOM artificial neural networks to extract valuable discriminative information of speech signal. Three types of secondary feature vectors were applied for phonemes classification. Spatial information, the magnitude component of each sample in spectro-temporal domain and the average of the amplitude components of each cluster points were considered as attributes in proposed features vectors. The proposed features were evaluated in compare to MDF, MFCC and WKM clustering-based features for phonemes classification. The experimental results indicate that the proposed features performed better for phoneme classification in comparison to MFCC, MDF and WKM-based features. The greatest improvement was obtained for back vowel 50.4. In consonants, the greatest improvement was achieved for unvoiced fricatives 24.3.

6. REFERENCES

1. Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A., "Mechanisms of noise robust representation of speech in primary auditory cortex", ***Proceedings of the National Academy of Sciences of the United States of America***, Vol. 111, No. 18, (2014), 6792–6797. doi:10.1073/pnas.1318017111
2. Patil, K., and Elhilali, M., "Biomimetic spectro-temporal features for music instrument recognition in isolated notes and solo phrases", ***Eurasip Journal on Audio, Speech, and Music Processing***, Vol. 2015, No. 1, (2015), 27. doi:10.1186/s13636-015-0070-9
3. Mesgarani, N., Slaney, M., and Shamma, S. A., "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations", ***IEEE Transactions on Audio, Speech and Language Processing***, Vol. 14, No. 3, (2006), 920–930. doi:10.1109/TSA.2005.858055
4. Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A., "Phoneme representation and classification in primary auditory cortex", ***The Journal of the Acoustical Society of America***, Vol. 123, No. 2, (2008), 899–909. doi:10.1121/1.2816572
5. Mesgarani, N., Fritz, J., and Shamma, S., "A computational model of rapid task-related plasticity of auditory cortical receptive fields", ***Journal of Computational Neuroscience***, Vol. 28, No. 1, (2010), 19–27. doi:10.1007/s10827-009-0181-3
6. Zulfiqar, I., Moerel, M., and Formisano, E., "Spectro-Temporal Processing in a Two-Stream Computational Model of Auditory Cortex", ***Frontiers in Computational Neuroscience***, Vol. 13, (2020), 1–18. doi:10.3389/fncom.2019.00095

7. Huang, C., and Rinzel, J., "A Neuronal Network Model for Pitch Selectivity and Representation", ***Frontiers in Computational Neuroscience***, Vol. 10, (2016), 1–17. doi:10.3389/fncom.2016.00057

8. De Martino, F., Moerel, M., Ugurbil, K., Goebel, R., Yacoub, E., and Formisano, E., "Frequency preference and attention effects across cortical depths in the human primary auditory cortex", ***Proceedings of the National Academy of Sciences of the United States of America***, Vol. 112, No. 52, (2015), 16036–16041. doi:10.1073/pnas.1507552112

9. Ruggles, D. R., Tausend, A. N., Shamma, S. A., and Oxenham, A. J., "Cortical markers of auditory stream segregation revealed for streaming based on tonotopy but not pitch", ***The Journal of the Acoustical Society of America***, Vol. 144, No. 4, (2018), 2424–2433. doi:10.1121/1.5065392

10. Valipour, S., Razzazi, F., Fard, A., and Esfandian, N., "A Gaussian clustering based voice activity detector for noisy environments using spectro-temporal domain", ***Signal Processing-An International Journal (SPIJ)***, Vol. 4, No. 4, (2010), 228–238

11. Yen, F. Z., Huang, M. C., and Chi, T.-S., "A two-stage singing voice separation algorithm using spectro-temporal modulation features", Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Vols 2015-January, (2015), 3321–3324.

12. Yin, P., Fritz, J. B., and Shamma, S. A., "Rapid spectrotemporal plasticity in primary auditory cortex during behavior", ***Journal of Neuroscience***, Vol. 34, No. 12, (2014), 4396–4408. doi:10.1523/JNEUROSCI.2799-13.2014

13. Lu, K., Liu, W., Zan, P., David, S. V., Fritz, J. B., and Shamma, S. A., "Implicit memory for complex sounds in higher auditory cortex of the ferret", ***Journal of Neuroscience***, Vol. 38, No. 46, (2018), 9955–9966. doi:10.1523/JNEUROSCI.2118-18.2018

14. Esfandian, N., Razzazi, F., and Behrad, A., "A clustering based feature selection method in spectro-temporal domain for speech recognition", ***Engineering Applications of Artificial Intelligence***, Vol. 25, No. 6, (2012), 1194–1202. doi:10.1016/j.engappai.2012.04.004

15. Esfandian, N., Razzazi, F., and Behrad, A., "A feature extraction method for speech recognition based on temporal tracking of clusters in spectro-temporal domain", AISP 2012 - 16th CSI International Symposium on Artificial Intelligence and Signal Processing, (2012), 12–17. doi:10.1109/AISP.2012.6313709

16. Esfandian, N., Razzazi, F., Behrad, A., and Valipour, S., "A feature selection method in spectro-temporal domain based on Gaussian Mixture Models", ***International Conference on Signal Processing Proceedings, ICSP***, (2010), 522–525. doi:10.1109/ICOSP.2010.5656082

17. Esfandian, N., "Phoneme classification using temporal tracking of speech clusters in spectro-temporal domain", ***International Journal of Engineering, Transactions A: Basics***, Vol. 33, No. 1, (2020), 105–111. doi:10.5829/ije.2020.33.01a.12

18.Nithya, A., Appathurai, A., Venkatadri, N., Ramji, D. R., and Anna Palagan, C., "Kidney disease detection and segmentation using artificial neural network and multi-kernel k-means clustering for ultrasound images", ***Measurement: Journal of the International Measurement Confederation***, Vol. 149, (2020), 106952. doi:10.1016/j.measurement.2019.106952

19. Peng, J., Wang, X., and Shang, X., "Combining gene ontology with deep neural networks to enhance the clustering of single cell RNA-Seq data", ***BMC Bioinformatics***, Vol. 20, No. 8, (2019), 1–12. doi:10.1186/s12859-019-2769-6

20. Nida, N., Irtaza, A., Javed, A., Yousaf, M. H., and Mahmood, M. T., "Melanoma lesion detection and segmentation using deep region based convolutional neural network and fuzzy C-means clustering", ***International Journal of Medical Informatics***, Vol. 124, (2019), 37–48. doi:10.1016/j.ijmedinf.2019.01.005

21. Hu, G., Wang, K., Peng, Y., Qiu, M., Shi, J., and Liu, L., "Deep Learning Methods for Underwater Target Feature Extraction and Recognition", ***Computational Intelligence and Neuroscience***, Vol. 2018, (2018). doi:10.1155/2018/1214301

22. Xia, Y., Braun, S., Reddy, C. K. A., Dubey, H., Cutler, R., and Tashev, I., "Weighted Speech Distortion Losses for Neural-Network-Based Real-Time Speech Enhancement", ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Vols 2020-May, (2020), 871–875. doi:10.1109/ICASSP40776.2020.9054254

23. Li, A., Yuan, M., Zheng, C., and Li, X., "Speech enhancement using progressive learning-based convolutional recurrent neural network", ***Applied Acoustics***, Vol. 166, (2020), 107347. doi:10.1016/j.apacoust.2020.107347

24. Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabí, M., Brown, K., Allopenna, P. D., Theodore, R. M., Monto, N., and Rueckl, J. G., "EARSHOT: A Minimal Neural Network Model of Incremental Human Speech Recognition", ***Cognitive Science***, Vol. 44, No. 4, (2020). doi:10.1111/cogs.12823

25. Garofolo, J., Lamel, L., Fiscus, J., and Pallett, D., DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus Documentatio, National Institute of Standards and Technology, Gaithersburg, MD, (1993).

26. Burges, C. J. C., "A tutorial on support vector machines for pattern recognition", ***Data Mining and Knowledge Discovery***, Vol. 2, No. 2, (1998), 121–167. doi:10.1023/A:1009715923555

27. Davis, S. B., and Mermelstein, P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", ***IEEE Transactions on Acoustics, Speech, and Signal Processing***, Vol. 28, No. 4, (1980), 357–366. doi:10.1109/TASSP.1980.1163420

28. Fartash, M., Setayeshi, S., and Razzazi, F., "A noise robust speech features extraction approach in multidimensional cortical representation using multilinear principal component analysis", ***International Journal of Speech Technology***, Vol. 18, No. 3, (2015), 351–365. doi:10.1007/s10772-015-9274-8

Persian Abstract

چکیده

در این مقاله، یک روش جدید استخراج ویژگی بر مبنای بازنمایی طیفی- زمانی سیگنال گفتار برای طبقه‌بندی واج ارائه شده است. در روش پیشنهادی، از شبکه عصبی برای خوشه‌بندی فضای طیفی- زمانی استفاده می‌شود. شبکه عصبی نگاشت ویژگی خود سامان (SOM) برای خوشه‌بندی فضای ویژگی‌ها اعمال شد. مقیاس، نرخ و فرکانس به عنوان اطلاعات مکانی هر نقطه و مولفه دامنه به عنوان ویژگی شباهت در الگوریتم خوشه‌بندی استفاده شد. سه مکانیزم برای انتخاب ویژگی‌ها در فضای طیفی- زمانی در نظر گرفته شد. اطلاعات مکانی خوشه‌ها، مولفه دامنه نقاط در فضای طیفی- زمانی و میانگین مولفه‌های دامنه نقاط هر خوشه به عنوان ویژگی‌های ثانویه در نظر گرفته شد. بردارهای ویژگی پیشنهادی برای طبقه‌بندی واج‌ها استفاده شد. نتایج نشان می‌دهد، بهبود قابل ملاحظه‌ای در نرخ طبقه‌بندی در دسته‌های مختلف واج‌ها در مقایسه با ویژگی‌های مبتنی بر خوشه‌بندی موجود به دست آمده است. نتایج به دست آمده با استفاده از ویژگی‌های جدید، آشکار می‌کند که خطای سیستم در کلیه زیرگروه‌های واج‌های صدادار و بی‌صدا در مقایسه با خوشه‌بندی K میانگین وزن‌دار بهبود یافته است.