# Use of Novel Feature Extraction Technique with Subspace Classifiers for Speech Recognition

**Conference Paper** · August 2007

DOI: 10.1109/PERSER.2007.4283894 · Source: IEEE Xplore



**2 authors:**

Serkan Gunal
Eskisehir Technical University
**42** PUBLICATIONS **1,248** CITATIONS

SEE PROFILE

Rifat Edizkan
Eskisehir Osmangazi University
**37** PUBLICATIONS **345** CITATIONS

SEE PROFILE

# Use of Novel Feature Extraction Technique with Subspace Classifiers for Speech Recognition

Serkan Gunal[1] and Rifat Edizkan[2]
[1]Anadolu University, Department of Computer Engineering, Eskisehir, Turkiye. serkangunal@anadolu.edu.tr
[2]Eskisehir Osmangazi University, Department of Electrical and Electronics Engineering, Eskisehir, Turkiye. redizkan@ogu.edu.tr

***Abstract*-Speech recognition is one of the fast moving research areas in pervasive services requiring human interaction. Like any type of pattern recognition system, selection of the feature extraction method and the classifier play a crucial role for speech recognition in terms of accuracy and speed. In this paper, an efficient wavelet based feature extraction method for speech data is presented. The feature vectors are then fed into three widely used linear subspace classifiers for recognition analysis. These classifiers are Class Featuring Information Compression (CLAFIC), Multiple Similarity Method (MSM) and Common Vector Approach (CVA). TI-DIGIT database is used to evaluate the performance of speaker independent isolated word recognition system designed. Experimental results indicate that the proposed feature extraction method together with the CLAFIC and CVA classifiers give considerably high recognition rates.**

## I. INTRODUCTION

Speech recognition is becoming a very important concept for any type of system requiring human interaction in today's hi-tech pervasive services. Controlling a system with speech rather than using hardware e.g. keyboard or keypad definitely much more easy and appealing. Performance of speech recognition system is however quite essential for the reliability of control. Recognition should be accurate and quick. Consequently, the feature extraction method and the classifier have a direct influence in speech recognition systems [1].

There exist several techniques of feature extraction for speech analysis. Linear predictive coding (LPC) and Mel Frequency Cepstral Coefficient (MFCC) are two widely used ones. LPC is based on a modelisation of the vocal tract. On the other hand, MFCC uses mel-scaled filterbanks inspired by the human ear scale [1].

Speech signal has a non-stationary structure but it is assumed that vocal tract is stationary for duration of 10 – 20 ms. Hence, speech signal is divided into small frames by a windowing process so that stationary operations can be performed. Then, the feature extraction is executed for each particular frame. Features extracted from each frame are then used together to represent the overall speech data. Frequency domain features are mostly used in speech analysis and they are obtained using Short Time Fourier Transform (STFT) [1].

Wavelet transform can be considered as an alternative to STFT. It has also certain advantages over STFT. The most important of all is that wavelets provide better time-frequency localization than STFT to be able to track sudden changes of speech signals [2].

In this paper, a wavelet based efficient feature extraction method is proposed. The success of the proposed method is evaluated on isolated speech recognition system using TI-DIGIT database. In the recognition system, subspace methods are used as pattern classifier. These methods are mostly aimed for classification rather than compression. The fundamental idea of the subspace methods is to find proper subspaces for particular classes within recognition system using covariance-correlation analysis [3]. The selected subspace classifiers for this paper are Class Featuring Information Compression (CLAFIC), Multiple Similarity Method (MSM) and Common Vector Approach (CVA). The individual performances of these three classifiers with the proposed feature extraction method are compared in the study.

In the forthcoming sections, the proposed wavelet based feature extraction method is presented; the subspace classifiers used in the study are explained; experimental work is described; and finally, the conclusion of the paper is given.

## II. FEATURE EXTRACTION

In this study, the feature extraction process for speech signals is handled using wavelet transforms. Wavelet transform is similar to Fourier transform in terms of breaking down signals into smaller parts for analysis where the Fourier transform uses sine waves of different frequencies while the wavelet transforms use wavelets, scaled and shifted versions of the "mother wavelet" for this job,

$$C(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi(\frac{t-b}{a})dt. \quad (1)$$

where $x$ is the signal to be transformed, $C$ is the continuous wavelet transform, $\Psi$ is the mother wavelet, $b$ is the shift and $a$ is the scale factor [2]. In Fourier transform, resolution is fixed at each scale. The wavelet transform, however, provides different resolution for each scale. This is the critical point that makes wavelet transform superior to Fourier due to better time-frequency localization.

Standard wavelet transform decomposes a signal (S) into its low (Approximation) and high frequency (Detail) components, in other words, its subbands. The approximation is then divided into a second-level approximation and detail level, and the process is repeated for further levels.

In wavelet packet analysis, however, the details as well as the approximations are decomposed to achieve full subband decomposition. Here, the decomposition is linear. In other words, each subband has the same bandwidth for respective level.

Since both low and high frequencies carry vital information for speech signals, wavelet packet analysis is much more suitable than standard wavelet transform. However, using a mel-scaled decomposition rather than linear scale may help more in terms of representation. Mel scaled filterbank is known to be best representation of human auditory system [4] and MFCCs computed using STFT, use this scale as it is apparent from the name. There also exist several studies indicating the efficiency of mel-scaled subband decomposition for speech data [5-6]. The feature extraction method, using mel-scaled wavelet transform is explained below.

i) Speech signal is first divided into frames to achieve a relatively stationary structure as mentioned before. Since utterances within the database have different length, using a fixed frame size would yield different number of frames for each utterance. To overcome this problem, the signal is divided into a fixed number of frames but with varying frame size. Determination of number of frames is here ad-hoc. Increasing the frame number would increase the resolution as well. However, increasing the number beyond certain threshold may not help to improve recognition performance at all. Instead, processing time for feature extraction would be excessive.

ii) For each frame, mel-scaled wavelet transform is applied. This decomposition produces 24 subbands. The frequency information of each subband is specified in Table I for the sampling frequency of 8 kHz.

TABLE I
SUBBAND FREQUENCY INFORMATION

| Subband | Frequency Range (Hz) | | | Bandwidth (Hz) |
|---|---|---|---|---|
| **1** | 0 | - | 62,5 | 62,5 |
| **2** | 62,5 | - | 125 | 62,5 |
| **3** | 125 | - | 187,5 | 62,5 |
| **4** | 187,5 | - | 250 | 62,5 |
| **5** | 250 | - | 312,5 | 62,5 |
| **6** | 312,5 | - | 375 | 62,5 |
| **7** | 375 | - | 437,5 | 62,5 |
| **8** | 437,5 | - | 500 | 62,5 |
| **9** | 500 | - | 562,5 | 62,5 |
| **10** | 562,5 | - | 625 | 62,5 |
| **11** | 625 | - | 687,5 | 62,5 |
| **12** | 687,5 | - | 750 | 62,5 |
| **13** | 750 | - | 875 | 125 |
| **14** | 875 | - | 1000 | 125 |
| **15** | 1000 | - | 1125 | 125 |
| **16** | 1125 | - | 1250 | 125 |
| **17** | 1250 | - | 1375 | 125 |
| **18** | 1375 | - | 1500 | 125 |
| **19** | 1500 | - | 1750 | 250 |
| **20** | 1750 | - | 2000 | 250 |
| **21** | 2000 | - | 2500 | 500 |
| **22** | 2500 | - | 3000 | 500 |
| **23** | 3000 | - | 3500 | 500 |
| **24** | 3500 | - | 4000 | 500 |

iii) Log energy of wavelet coefficients within each of 24 subbands is then calculated. Thus, 24 scalar values are obtained per frame.

iv) Finally, the scalar values from all available frames are combined so that the feature vector for respective speech signal is composed.

Following the above procedure, the resulting feature vectors are ready to be fed into the subspace classifiers mentioned before.

## III. SUBSPACE CLASSIFIERS

The idea underlying the subspace classifiers originates from compression and optimal reconstruction of multidimensional data with linear principal components. The use of linear subspaces as class models is based on the assumption that the vector distribution in each class lies approximately on a lower-dimensional subspace of the feature space. A test vector from an unknown class can be classified by computing the shortest distance among all subspaces, each one representing single class [3].

Subspace methods mostly aimed for classification rather than compression. The fundamental idea of the learning methods is to modify the subspace bases so that minimum misclassification error is achieved. In this work, three different subspace classifiers are used in the recognition process: CLAFIC, MSM and CVA.

### *A. CLAFIC*

One of the most widely used subspace methods is CLAFIC algorithm introduced by Watanabe et al. [7]. CLAFIC simply forms the base matrices for the classifier subspaces from the eigenvectors of the class-conditional correlation matrices. For each class $c$, the correlation matrix $R_c$ is estimated. Certain number of eigenvectors corresponding to the largest eigenvalues of $R_c$ is then used as the columns of the basis matrix $U_c$.

The sample mean $\mu$ of the pooled training set may be subtracted from the pattern vectors before they are classified or used in initializing CLAFIC classifier. Since the class-conditional correlations $R_c$ of the input vectors $x$ differ from the corresponding class-wise covariances $\sum_c$, the first eigendirection in each class merely reflects the direction of the class mean from the pooled mean translated to the origin.

### *B. MSM*

It is explained in CLAFIC that certain number of eigenvectors in the order of decreasing eigenvalues constitutes the basis matrix for respected class. Selection of this number is ad-hoc and directly affects the recognition performance. Iijima et al. [8] have selected to weight each basis vector with the corresponding eigenvalue in their MSM classifier as an alternative to ad-hoc selection of CLAFIC algorithm. This method emphasizes the effect of the most prominent directions for which $\lambda_{ic} / \lambda_{1c} \cong 1$, where $i$ is vector dimension and $c$ is the class index. Since the influence of the relatively less important eigenvectors that have multipliers $\lambda_{ic} / \lambda_{1c} \cong 0$ diminishes gradually, the selection of the subspace dimension is therefore less essential unlike CLAFIC.

*C. CVA*

CVA is another subspace-based classifier developed in recent years. It has been successfully used in speech and some other pattern recognition applications [9-12]. In CVA, a unique common vector that represents common or invariant features of a class is obtained. CVA has been applied to two cases: sufficient data ($m \geq n$) and insufficient data ($m < n$), where $m$ and $n$ represents the number of vectors in training set and the number of elements in feature vector respectively.

In CVA, the feature space is divided into two orthogonal subspaces: the difference subspace ($B$) and the indifference subspace ($B^{\perp}$). Eigenvectors that span the difference and indifference subspaces are obtained from the within-class covariance matrix of a class. Let $a_i^c$ represent a feature vector in class $c$. In the sufficient data case, the common vector and the subspace division are obtained from the minimization of the following criterion:

$$F^c = \sum_{i=1}^{m} \left\| P_c^{\perp}(a_i^c - a_{ave}^c) \right\|^2. \quad (2)$$

where $P^{c\perp}$ is the projection matrix for the indifference subspace and $a_{ave}^c$ is the mean vector of respective class. The criterion in (2) is minimized for each class separately. The criterion states that the features of a class should be close to the class average within the indifference subspace. A solution to the optimization problem of (2) is to solve the generalized eigenvalue problem in (3) where $\Phi^c$ is within-covariance matrix, $\lambda_j^c$ and $u_j^c$ denote eigenvalue - eigenvector pair of the covariance matrix for class $c$ respectively.

$$(\Phi^c - \lambda_j^c I)u_j^c = \vec{0} \qquad j = 1,2,..,m. \quad (3)$$

For sufficient data case ($m \geq n$), to minimize $F^c$, the indifference subspace should be constructed from the eigenvectors corresponding to the smallest eigenvalues [12]. Let the eigenvalues of the within-class covariance matrix $\Phi^c$ be sorted in ascending order. Here, the eigenvectors corresponding to the first $k$ smallest eigenvalues span the indifference subspace. The rest of eigenvectors ($n-k$) will then span the difference subspace. The projection matrix for indifference subspace is then computed as in (4).

$$P^{c^{\perp}} = \sum_{j=1}^{k} u_j^c u_j^{c^T}. \quad (4)$$

In sufficient case, the common vector of a class is defined as

$$a_{com}^c = P^{c^{\perp}} a_{ave}^c. \quad (5)$$

In CVA, minimum Euclidean distance is used for classification.

$$c^* = \arg\min_{1 \leq c \leq K} \left\| P^{c^{\perp}} a_x - a_{com}^c \right\|. \quad (6)$$

According to (6), an unknown feature $a_x$ is assigned to class $c$ if the minimum distance is obtained for this class among all.

## IV. EXPERIMENTAL WORK

In the experimental study, during the feature extraction step, Daubechies-32 mother wavelet is used for wavelet transform operations. Additionally, different frame numbers (1, 2, 4, and 10) are selected for performance comparison. These numbers correspond to 24, 48, 96 and 240 dimensional feature vectors keeping in mind that a single frame is represented by 24 scalar values as mentioned before in the feature extraction section.

Eigenvalues obtained from the correlation and within-class covariance matrix of 96 dimensional feature vectors belonging to class "0" are shown in Fig.1 and 2 respectively. Based on these variations of the eigenvalues, CLAFIC classifier uses a few of the eigenvectors corresponding to the largest eigenvalues given in Fig. 1 to compose the basis matrix. Conversely, CVA uses most of the eigenvectors corresponding to the smallest eigenvalues in Fig. 2 to setup the indifference subspace projection matrix of the regarding class. MSM however does not need a selection of eigenvectors as indicated. It instead uses all of them but with weighting.

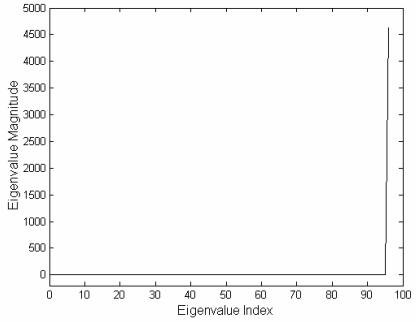


Fig. 1. Eigenvalues obtained from the correlation matrix of 96 dimensional feature vectors belonging to class "0".

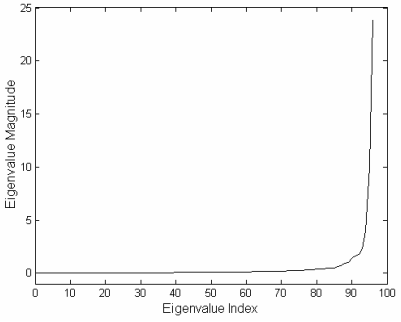


Fig. 2. Eigenvalues obtained from the within-class covariance matrix of 96 dimensional feature vectors belonging to class "0".

The performance of the designed speech recognition system is investigated and evaluated on the TI-DIGIT database sampled at 8 kHz in this paper. Among a total of 450 utterances by both male and female speakers for each of 10 digits available in the database, leave-50-out method is applied for 9 different combinations so that 400 samples are used for training while the remaining 50 samples are reserved for testing in each combination. By this way, fair performance evaluation of the designed recognition system is carried out.

The recognition results of three subspace classifiers, CLAFIC, MSM and CVA, using the feature vectors with different number of frames are shown in Table II, III and IV respectively for 10 classes of TI-DIGIT database.

TABLE II
RECOGNITION RESULTS (%) FOR CLAFIC CLASSIFIER

| Digit | 1-frame | 2-frame | 4-frame | 10-frame |
|---|---|---|---|---|
| 0 | 62.89 | 89.33 | 97.56 | 99.33 |
| 1 | 50.89 | 76.22 | 98.00 | 99.56 |
| 2 | 66.89 | 84.89 | 98.67 | 98.00 |
| 3 | 72.00 | 88.00 | 98.44 | 98.89 |
| 4 | 83.78 | 91.11 | 97.11 | 98.67 |
| 5 | 76.67 | 77.78 | 95.11 | 98.67 |
| 6 | 60.44 | 90.00 | 95.78 | 99.11 |
| 7 | 80.22 | 86.00 | 95.78 | 99.33 |
| 8 | 81.11 | 90.22 | 95.78 | 97.78 |
| 9 | 61.33 | 70.89 | 92.22 | 96.67 |
| Average | **69.62** | **84.44** | **96.44** | **98.60** |

TABLE III
RECOGNITION RESULTS (%) FOR MSM CLASSIFIER

| Digit | 1-frame | 2-frame | 4-frame | 10-frame |
|---|---|---|---|---|
| 0 | 56.22 | 84.00 | 90.89 | 91.11 |
| 1 | 60.44 | 61.78 | 84.22 | 88.00 |
| 2 | 66.89 | 90.67 | 81.56 | 87.56 |
| 3 | 68.44 | 80.00 | 89.78 | 91.33 |
| 4 | 88.44 | 93.11 | 94.67 | 92.67 |
| 5 | 80.44 | 76.67 | 87.11 | 91.11 |
| 6 | 69.33 | 87.33 | 90.89 | 85.11 |
| 7 | 71.11 | 84.67 | 71.11 | 68.44 |
| 8 | 71.78 | 62.22 | 78.00 | 82.89 |
| 9 | 70.00 | 67.56 | 76.67 | 88.44 |
| Average | **70.31** | **78.80** | **84.49** | **86.67** |

TABLE IV
RECOGNITION RESULTS (%) FOR CVA CLASSIFIER

| Digit | 1-frame | 2-frame | 4-frame | 10-frame |
|---|---|---|---|---|
| 0 | 72.67 | 91.56 | 98.22 | 99.78 |
| 1 | 79.11 | 68.22 | 97.56 | 99.33 |
| 2 | 84.22 | 92.67 | 98.22 | 98.00 |
| 3 | 71.33 | 96.89 | 98.67 | 98.89 |
| 4 | 93.11 | 95.33 | 98.22 | 98.44 |
| 5 | 82.44 | 80.67 | 93.11 | 98.89 |
| 6 | 80.22 | 94.89 | 95.11 | 98.67 |
| 7 | 80.00 | 94.67 | 97.78 | 99.11 |
| 8 | 85.56 | 88.22 | 96.89 | 97.56 |
| 9 | 74.89 | 81.78 | 92.22 | 97.33 |
| Average | **80.36** | **88.49** | **96.60** | **98.60** |

According to the results, CVA classifier provides the best recognition rate in each case for all classes tested. CLAFIC is the runner-up while MSM classifier is the third. The results reveal that 1- or 2-frame resolution is not enough to get satisfactory recognition accuracy in any of the classifiers. However, 4-frame corresponding to 96-dimensional feature seems to be the optimal value. It gives approximately 97% correct recognition rate for both CLAFIC and CVA classifier. It is also obvious that selecting a frame number beyond 4 offers relatively low recognition performance improvement where the overall processing time would increase much more.

## V. CONCLUSION

In this paper, we presented a novel feature extraction technique for speech recognition based on mel-scaled wavelet transforms. To overcome varying utterance length issue, fixed number of frames with varying frame size methodology is used.

The features extracted with this method are then fed into three different linear subspace classifiers for performance comparison at different resolutions. The subspace methods are widely used for classification tasks rather than compression. These methods use correlation-covariance analysis and find a proper subspace projection for particular classes.

In conclusion, it is obvious from the recognition results that the proposed feature extraction method together with the subspace classifiers is well suited for speech recognition systems used in pervasive services. Wavelet based features offer successful representation of speech data with relatively low dimension. The linear subspace classifiers using those features have less algorithmic complexity and give reasonable recognition rates as well.

## REFERENCES


[1] L. Rabiner, BH Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[2] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, 2001.

[3] E. Oja, *Subspace Methods of Pattern Recognition*, Research Studies Press, 1983.

[4] S.B. Davis, P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. on Acoustic, Speech and Signal Processing, vol. 28, pp.357–366, 1980.

[5] O. Farooq and S. Datta, "Mel filter-like admissible wavelet packet structure for speech recognition", IEEE Signal Processing Letters, vol. 8, pp. 196–198, 2001.

[6] S. Gunal, R. Edizkan, "Wavelet based discriminative feature extraction for speech recognition", The Proc. of International Conference on Modeling and Simulation, Konya, Turkiye, pp. 621–624, 2006.

[7] S. Watanabe, P. F. Lambert, C. A. Kulikowski, J. L. Buxton, R. Walker, "Evaluation and selection of variables in pattern recognition", In J. Tou (Ed.), Computer and Information Sciences II. New York: Academic Press, 1967.

[8] T. Iijima, H. Genchi, K. Mori, "A theory of character recognition by pattern matching method", The Proc. of the 1st International Joint Conference on Pattern Recognition, Washington, DC, pp. 50–56, 1973.

[9] M. B. Gülmezoğlu, V. Dzhafarov, M. Keskin, A. Barkana, "A novel approach to isolated word recognition", IEEE Trans. on Speech and Audio Processing, vol. 7 (6), pp. 620–628, 1999.

[10] M. B. Gülmezoğlu, V. Dzhafarov, A. Barkana, "The common vector approach and its relation to the principal component analysis", IEEE Trans. on Speech and Audio Processing, vol. 9 (6), pp. 655–662, 2001.

[11] S. Gunal, S. Ergin., M. B. Gülmezoğlu, Ö. N. Gerek, "On feature extraction for spam e-mail detection", LNCS, vol. 4105, pp. 635–642, 2006.

[12] M. B. Gülmezoğlu, V. Dzhafarov, R. Edizkan, A. Barkana, "The common vector approach and its comparison with other subspace methods in case of sufficient data", Computer Speech and Language, vol. 21 (2), pp.266–281, 2007.