Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 56382, 17 pages
doi:10.1155/2007/56382

*Research Article*

# A New Method to Represent Speech Signals Via Predefined Signature and Envelope Sequences

**Ümit Güz,[1,2] Hakan Gürkan,[1] and Binboga Sıddık Yarman[3,4]**

[1] *Department of Electronics Engineering, Engineering Faculty, Işık University, Kumbaba Mevkii, Şile, 34980 Istanbul, Turkey*
[2] *SRI-International, Speech Technology and Research (STAR) Laboratory, 333 Ravenswood Avenue, Menlo Park, CA 94025, USA*
[3] *Department of Electrical-Electronics Engineering, College of Engineering, Istanbul University, Avcılar, 34230 Istanbul, Turkey*
[4] *Department of Physical Electronics, Graduate School of Science and Technology, Tokyo Institute of Technology, (Ookayama Campus) 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan*

Received 3 June 2005; Revised 28 March 2006; Accepted 30 April 2006

Recommended by Kostas Berberidis

A novel systematic procedure referred to as "SYMPES" to model speech signals is introduced. The structure of SYMPES is based on the creation of the so-called predefined "signature $S = \{S_R(n)\}$ and envelope $E = \{E_K(n)\}$" sets. These sets are speaker and language independent. Once the speech signals are divided into frames with selected lengths, then each frame sequence $X_i(n)$ is reconstructed by means of the mathematical form $X_i(n) = C_i E_K(n) S_R(n)$. In this representation, $C_i$ is called the gain factor, $S_R(n)$ and $E_K(n)$ are properly assigned from the predefined signature and envelope sets, respectively. Examples are given to exhibit the implementation of SYMPES. It is shown that for the same compression ratio or better, SYMPES yields considerably better speech quality over the commercially available coders such as G.726 (ADPCM) at 16 kbps and voice excited LPC-10E (FS1015) at 2.4 kbps.



## 1. INTRODUCTION

Transmission and storage of speech signals are widespread in modern communications systems. The field of speech representation or compression is dedicated to finding new and more efficient ways to reduce transmission bandwidth or storage area while maintaining high quality of hearing [1].

In the past, a number of new algorithms based on the use of numerical, mathematical, statistical, and heuristic methodologies were proposed in order to represent, code, or compress the speech signals. For example, in the construction of speech signals, linear predictive coding (LPC) techniques such as LPC-10E (FS1015) utilize low bit rates at 2.4 kbps with acceptable hearing quality. Pulse code modulation (PCM) techniques such as G.726 (ADPCM) yield much better hearing quality over LPC-10E but demand higher bit rates of 32 or 16 kbps [1–3].

In our previous work [4–7], efficient methods to model speech signals with low bit rates and acceptable hearing quality were introduced. In these methods, one would first examine the signals in terms of their physical features, and then find some specific waveforms to best describe the signals, called signature functions. Signature functions of speech signals are obtained by using energy compaction property of the principal component analysis (PCA) [8–14]. PCA also provides optimal solution via minimization of the error in the least mean square (LMS) sense. The new method presented in this paper significantly improves the results of [4–7] by introducing the concept of "signal envelope" in the representation of speech signals. Thus, the new mathematical form of the frame signal $X_i$ is proposed as $X_i \approx C_i E_K S_R$ where $C_i$ is a real constant called the gain factor, $S_R$ and $E_K$ are properly extracted from the so-called predefined signature set $S = \{S_R\}$ and predefined envelope set $E = \{E_K\}$ or in short PSS and PES, respectively. It is exhibited that PSS and PES which are generated as the result of this work are independent of the speaker and the language spoken. It is also worth mentioning that if the proposed modeling technique is employed in communication, it results in substantial reductions in transmission bandwidth. If it is used for digital recording, it provides great savings in the storage area. In the following sections theoretical aspects of the proposed modeling technique are presented and the implementation details are discussed. Implementation results are summarized. Possible applications and directions for future research are included in the conclusion. It is noted that the initial results of the new method were introduced in [15–17]. In this paper however, results of [15–17] are considerably enhanced by creating almost complete PSS and PES for different languages utilizing the *Phonetics Handbook* prepared by the International Phonetics Association (IPA) [18].

## 2. THE PROPOSED METHOD

It would be appropriate to extract the statistical features of the speech signals over a reasonable length of time. For the sake of practicality, we present the new technique on the discrete time domain since all the recordings are made with digital equipment. Let $X(n)$ be the discrete time domain representation of a recorded speech piece with $N$ samples.

Let this piece be analyzed frame by frame. In this representation, $X_i(n)$ denotes a selected frame as shown in Figure 1. Then, the following main statement and the related definitions are proposed which constitute the basis of the new modeling technique.

### 2.1. Main statement

Referring to Figure 1, for any time frame $i$, the sampled speech signal which is given by the vector $X_i$ of length $L_F$ can be approximated as

$$X_i \cong C_i E_K S_R, \tag{1}$$

where

(i) $C_i$ is a real constant and it is called the gain factor,
(ii) $K$, $R$, $N_E$, and $N_S$ are integers such that $K \in \{1,2,\ldots,N_E\}$, $R \in \{1,2,\ldots,N_S\}$,
(iii) the signature vector $S_R^T = [s_{R1} \ s_{R2} \ \ldots \ s_{RL_F}]$ is generated utilizing the statistical behavior of the speech signals and the term $C_iS_R$contains almost full energy of $X_i$ in the LMS sense,
(iv) $E_K$ is ($L_F$ by $L_F$) diagonal matrix such that

$$E_K = \begin{bmatrix} e_{K1} & 0 & 0 & \ldots & 0 \\ 0 & e_{K2} & 0 & \ldots & 0 \\ 0 & 0 & e_{K3} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & e_{KL_F} \end{bmatrix} \tag{2}$$

and acts as an envelope term on the quantity $C_iS_R$ which also reflects the statistical properties of the speech signal under consideration,
(v) the integer $L_F$ designates the total number of samples in the $i$th frame.

Now, let us verify the main statement.

### 2.2. Verification of the main statement

The sampled speech signal sequence $x(n)$ can be written as

$$x(n) = \sum_{i=1}^{N} x_i \delta_i(n-i). \tag{3}$$

In (3), $\delta_i(n)$ represents the unit sample; $x_i$ designates the measured value of the sequence $x(n)$ at the $i$th sample. $x(n)$ can also be expressed in vector form as

$$X^T = \begin{bmatrix} x(1) & x(2) & \ldots & x(N) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \ldots & x_N \end{bmatrix}. \tag{4}$$

In this representation, $X$ is called the main frame vector (MFV) and it may be divided into frames with equal lengths, having, for example, 16, 24, 32, 64, or 128 samples and so forth. In this case, MFV which is also designated by $M_F$ is obtained by means of the frame vectors $\{X_1, X_2, \ldots, X_{NF}\}$

$$M_F = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{N_F} \end{bmatrix}, \qquad M_F^T = \begin{bmatrix} X_1^T & X_2^T & \ldots & X_{N_F}^T \end{bmatrix}, \tag{5}$$

where

$$X_i = \begin{bmatrix} x_{(i-1)L_F+1} \\ x_{(i-1)L_F+2} \\ \vdots \\ x_{iL_F} \end{bmatrix}, \quad i = 1,2,\ldots,N_F. \tag{6}$$

$N_F = N/L_F$ denotes the total number of frames in $X$. Obviously, integers $N$ and $L_F$ must be selected in such a way that $N_F$ also becomes an integer.

As it is given by [7], each frame sequence or vector $X_i$ can be spanned in a vector space formed by the orthonormal vectors[1] $\{\phi_{ik}\}$ such that

$$X_i = \sum_{k=1}^{L_F} c_k \phi_{ik}, \quad k = 1,2,\ldots,L_F, \tag{7}$$

where the frame coefficients $c_k$ are obtained as

$$c_k = \phi_{ik}^T X_i, \quad k = 1,2,\ldots,L_F \tag{8}$$

and $\{\phi_{ik}\}$ are generated as the eigenvectors of the frame correlation matrix $R_i$

$$R_i = E[X_iX_i^T] = \begin{bmatrix} r_i(1) & r_i(2) & r_i(3) & \ldots & r_i(L_F) \\ r_i(2) & r_i(1) & r_i(2) & \ldots & r_i(L_F-1) \\ r_i(3) & r_i(2) & r_i(1) & \ldots & r_i(L_F-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_i(L_F) & r_i(L_F-1) & r_i(L_F-2) & \ldots & r_i(1) \end{bmatrix} \tag{9}$$

constructed with the entries;

$$r_i(d+1) = \frac{1}{L_F} \sum_{j=[(i-1)\cdot L_F+1]}^{[i\cdot L_F-d]} x_j x_{j+d}, \quad d = 0,1,2,\ldots,L_F-1. \tag{10}$$

[1] It is noted that orthonormal vector $\phi_{ik}$ satisfies $\phi_{ik}^T\phi_{ik} = 1$.

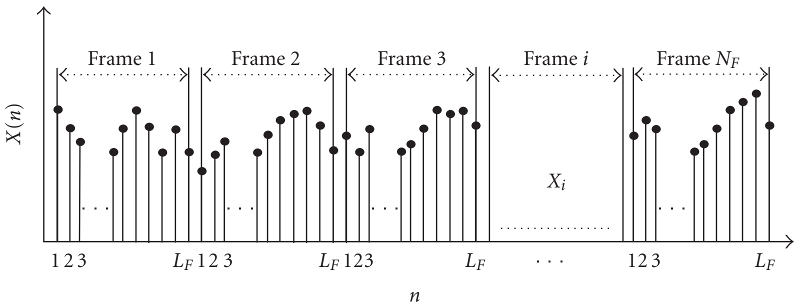


FIGURE 1: Segmentation of speech signals frame by frame.

In (9) $E[\cdot]$ designates the expected value of a random variable. Obviously, $R_i$ is real, symmetric, positive semidefinite, and Toeplitz which in turn yields real, distinct, and nonnegative eigenvalues $\lambda_{ik}$ satisfying the relation $R_i\phi_{ik} = \lambda_{ik}\phi_{ik}$. Let the eigenvalues be sorted in descending order such that $(\lambda_{i1} \geq \lambda_{i2} \geq \lambda_{i3} \geq \cdots \geq \lambda_{iL_F})$ with corresponding eigenvectors $\{\phi_{ik}\}$. Then, the total energy of the frame $i$ is given by $X_i^T X_i$:

$$X_i^T X_i = \sum_{k=1}^{L_F} x_{ik}^2 = \sum_{k=1}^{L_F} c_{ik}^2. \tag{11a}$$

In the mean time, the expected value of this energy is expressed as

$$E\left[\sum_{k=1}^{L_F}\left[c_{ik}^2\right]\right] = \sum_{k=1}^{L_F} \phi_{ik}^T E\left[\left(X_i X_i^T\right)\right]\phi_{ik} = \sum_{k=1}^{L_F} \phi_{ik}^T R_i \phi_{ik} = \sum_{k=1}^{L_F} \lambda_{ik}. \tag{11b}$$

In (11), contributions of the higher order terms become negligible, perhaps after $p$ terms. In this case, (7) may be truncated. The simplest form of (7) is obtained by setting $p = 1$.

As an example, let us consider a randomly selected 16 sequential voice frames formed with $L_F = 16$ samples. In this case, one would end up with 16 distinct positive-real eigenvalues in descending order for each frame. If one plots all the eigenvalues on a frame basis then, Figure 2 follows. This figure shows that the eigenvalues become drastically smaller after the first one. Moreover, if one varies the frame length $L_F$ as a parameter to further reduce the effect of the second- and higher-order terms then, almost full energy of the signal frame is captured within the first term of (7). Hence,

$$X_i \cong c_1 \phi_{i1}. \tag{12}$$

That is why $\phi_{i1}$ is called the signature vector since it contains most of the useful information of the original speech frame under consideration. Once (12) is obtained, it can be converted to an equality by means of an envelope term $E_i$ which is a diagonal matrix for each frame. Thus, $X_i$ is computed as

$$X_i = C_i E_i \phi_{i1}. \tag{13}$$

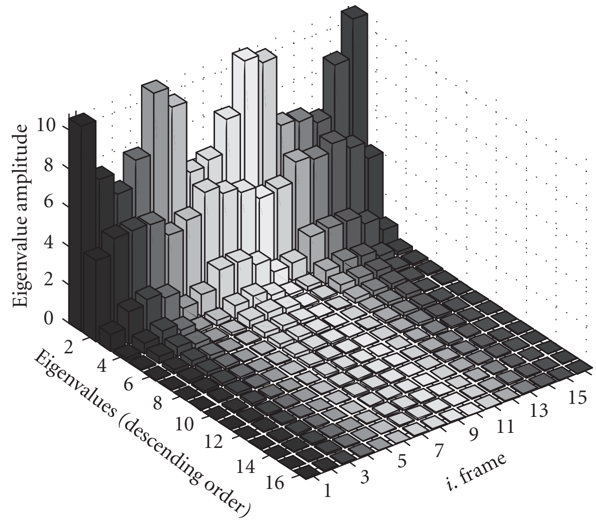


FIGURE 2: Plot of the 16 distinct eigenvalues in a descending order for 16 adjacent speech frames.

In (13), diagonal entries $e_{ir}$ of the matrix $E_i$ are determined in terms of the entries of $\phi_{i1}^T = [\phi_{i11} \ \cdots \ \phi_{i1r} \ \cdots \ \phi_{i1L_F}]$ and $X_i^T = [x_{i1} \ \cdots \ x_{ir} \ \cdots \ x_{iL_F}]$ by simple division.

$$e_{ir} = \frac{x_{ir}}{C_i \phi_{i1r}}, \quad (r = 1, 2, \ldots, L_F). \tag{14}$$

In essence, the quantities $e_{ir}$ of (14) somewhat absorb the remaining energy of the terms eliminated by truncation process of (7). This approach constitutes the basis of the new speech modeling technique as follows.

In this research, several tens of thousands of speech pieces were investigated frame by frame and several thousands of "signature and envelope sequences" were generated. It was observed that patterns obtained by plotting the envelope $e_i(n)$ ($e_{ir}$ versus *frame index-n* $= 1, 2, \ldots, L_F$) and signature sequences $\phi_{i1}(n)$ ($\phi_{i1r}$ versus *frame index-n* $= 1, 2, \ldots, L_F$) exhibit similarities. Some of these patterns are shown in Figures 3 and 4, respectively. It is deduced that these similar patterns are obtained due to the quasistationery behavior of the speech signals. In this case, one can eliminate the similar patterns and thus, constitute the so-called "predefined signature sequence" and "predefined envelope sequence" sets

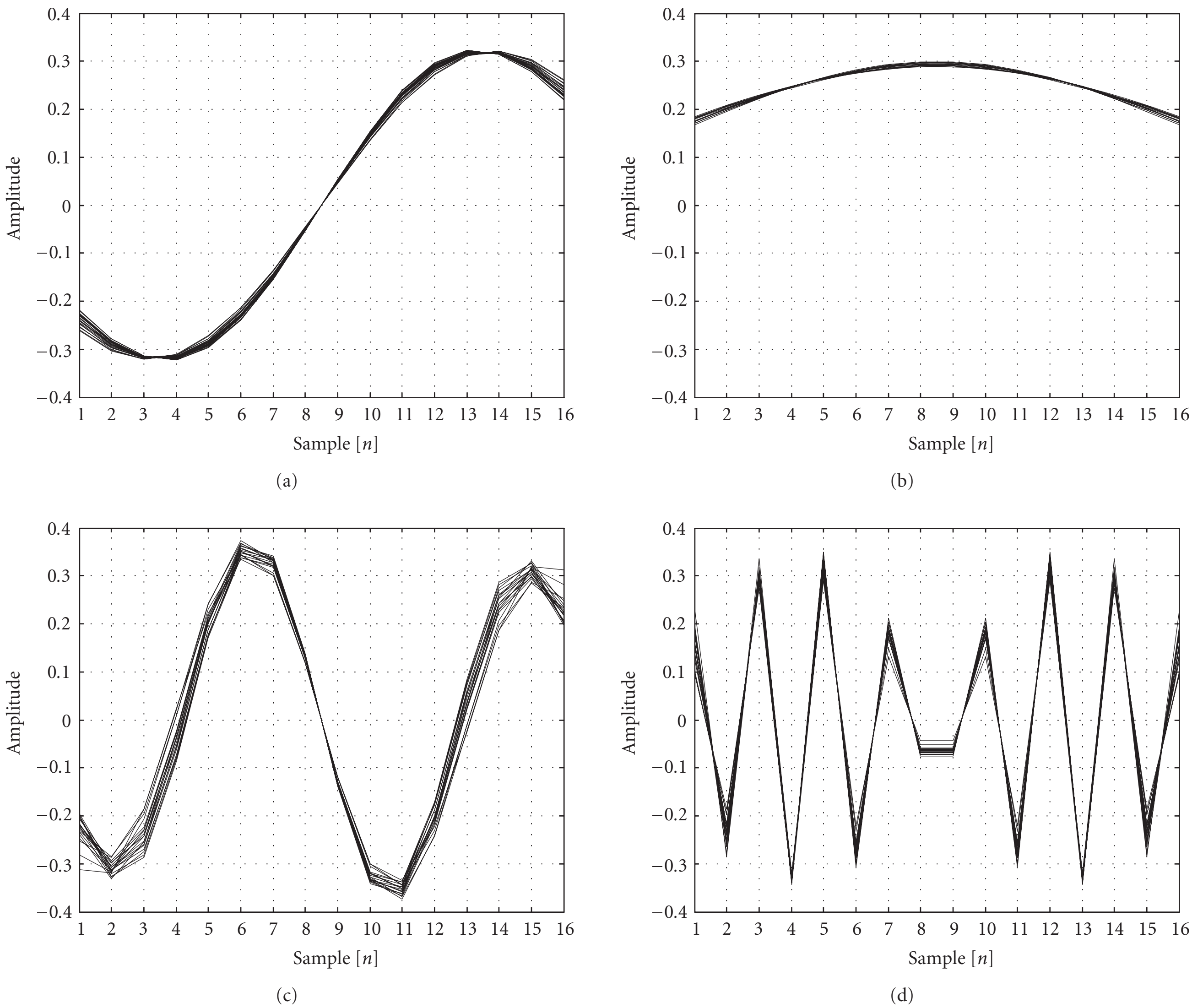


Figure 3: Some selected eigenvectors which exhibit similar patterns ($L_F = 16$).

constructed with one of a kind, or unique patterns. All the above groundwork leads one to propose "a novel systematic procedure to model speech signals by means of PSS and PES." In short, the new numerical procedure is called "SYMPES" and it is outlined in the following section.

### 2.3. A novel systematic procedure to model speech signals via predefined envelope and signature sets: SYMPES

SYMPES is a systematic procedure to model speech signals in four major steps described as follows.

*Step 1.* Selection of speech pieces to create signature and envelope sequences.

(i) For a selected frame length $L_F$, investigate variety of speech pieces frame by frame which describe the major characteristics of speakers and languages to determine signature and envelope sequences. This step may result in hundreds of thousand of signature and envelope sequences for different languages. However, these sequences exhibit too many similar patterns subject to elimination.

*Step 2.* Elimination of similar patterns.

(i) Eliminate the similar patterns of signature and envelope sequences to end up with unique shapes. Then, form the PSS and PES utilizing the unique patterns.

*Step 3.* Reconstruction of speech frame by frame.

(i) Once PSS and PES are formed, one is ready to synthesize a given speech piece $X(n)$ of length $N$ frame by frame. In this case, divide $X(n)$ into frames of length $L_F$ in a sequential manner to form the MFV of (5). Then, for each frame $X_i$, find the best approximation $X_{Ai} = C_i E_K S_R$ by computing the real coefficient $C_i$,

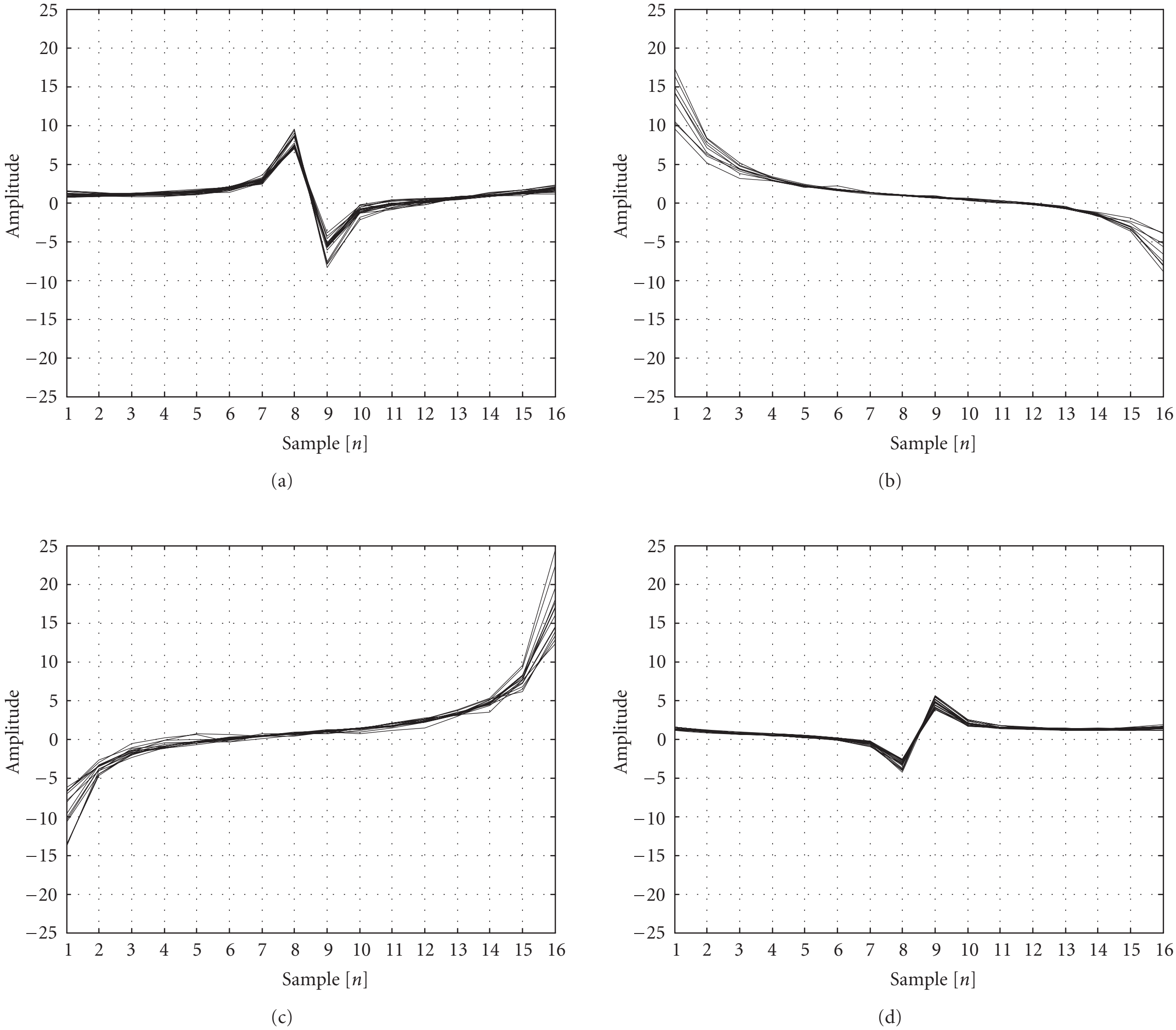


Figure 4: Some selected envelope vectors which exhibit similar patterns ($L_F = 16$).

pulling $E_K$ from PES and $S_R$ from PSS to minimize the frame error defined by $\varepsilon_i(n) = X_i(n) - C_i E_K S_R$, in the LMS sense.

(ii) Eventually, sequences $X_{Ai}$ are collected under the approximated main frame vector

$$M_{AF} = \begin{bmatrix} X_{A1} \\ X_{A2} \\ \vdots \\ X_{AN_F} \end{bmatrix} \text{ to reconstruct the speech as} \tag{15}$$

$$X_A(n) = \{X_{A1}, X_{A2}, \ldots, X_{AN_F};\ N_F = N/NL_F\} \approx X(n).$$

*Step 4.* Elimination of the background noise due to the reconstruction process by using a moving average post-filter.

(i) At the end of the third step, the reconstructed signal may contain unexpected spikes in merging process of the speech frames in sequential order. These spikes may cause unexpected background noise which may be classified as the musical noise. It was experienced that the musical noise can significantly be reduced by means of a moving average post-filter. In this regard, one may utilize a simple moving average finite impulse response filter. Nevertheless, an optimum filter can be selected by trial and error depending on the environmental noise, and the operational conditions.

In the following section, an elimination process of similar patterns of signature and envelope sequences are described [19]. At this point, it should be noted that the modeler is free to employ any other elimination or vector reduction technique to enhance the quality of hearing. In this regard, one may even wish to utilize the LBG vector quantization technique with different varieties to reduce the signature and the envelope sets as desired [20]. Essentials of the sample selection to generate PSS and PES are introduced in Section 4. Computational details to construct PSS and PES are presented by Algorithm 1. The numerical aspects of the speech reconstruction process are given by Algorithm 2.

### 2.4. Elimination of similar patterns

One of the useful tools to measure the similarities between two sequences is known as the Pearson correlation coefficient (PCC). PCC is designated by $\rho_{YZ}$ and given as [19]

$$\rho_{YZ} = \frac{\sum_{i=1}^{L}(y_i z_i) - \left[\sum_{i=1}^{L} y_i \sum_{i=1}^{L} z_i\right]/L}{\sqrt{\left[\sum_{i=1}^{L} y_i^2 - \left(\sum_{i=1}^{L} y_i\right)^2/L\right]\left[\sum_{i=1}^{L} z_i^2 - \left(\sum_{i=1}^{L} z_i\right)^2/L\right]}}. \tag{16}$$

In the above formula $Y = [y_1\ y_2\ \dots\ y_L]$ and $Z = [z_1\ z_2\ \dots\ z_L]$ are two sequences subject to comparison. Clearly, (16) indicates that $\rho_{YZ}$ is always between $-1$ and $+1$. $\rho_{YZ} = 1$ indicates that two vectors are identical. $\rho_{YZ} = 0$ corresponds to completely uncorrelated vectors. On the other hand, $\rho_{YZ} = -1$ refers to perfectly opposite pair of vectors (i.e., $Y = -Z$). For the sake of practicality, it is assumed that the two sequences are almost identical if $0.9 \le \rho_{YZ} \le 1$. Hence, similar patterns of signature and envelope sequences are eliminated accordingly. Thus, the signature vectors which have unique patterns are combined under the set called predefined signature set PSS $= \{S_{n_s}(n);\ n_s = 1, 2, \dots, N_S\}$. The integer $N_S$ designates the total number of elements in this set. Similarly, reduced envelope sequences are combined under the set called predefined envelope set PES $= \{E_{n_e}(n);\ n_e = 1, 2, \dots, N_E\}$. The integer $N_E$ designates the total number of unique envelope sequences in PES. At this point, it should be noted that members of PSS are not orthogonal. They are just the unique patterns of the first eigenvectors of various speech frames obtained from thousands of different experiments. In Figures 5 and 6, some selected one of a kind signature and envelope sequences are plotted point by point against their entry indices resulting in the signature and envelope patterns, respectively.

All of the above explanations endorse the phrasing of the main statement that any speech frame $X_i$ can be modeled in terms of the gain factor $C_i$, predefined signature $S_R$, and envelope $E_K$ terms as $X_i \approx C_i E_K S_R$. In the following section, algorithms are summarized to generate PSS and PES.

## 3. GENERATION OF PSS AND PES AND THE RECONSTRUCTION PROCESS OF SPEECH

The heart of the newly proposed method to model speech signals is based on the generation of the PSS and PES. Therefore, in this section first an algorithm is outlined to construct PSS and PES (Algorithm 1) then, synthesis or reconstruction process of speech signals is detailed (Algorithm 2).

### 3.1. Algorithm 1: generation of the predefined signature and envelope sets

#### Inputs

(i) Main frame sequence of the speech piece $\{X(n), n = 1, 2, \dots, N\}$.
Herewith, sample speech pieces given by the IPA Handbook were utilized [18]. This handbook includes phonetics properties (vowels, consonants, tones, stress, conventions, etc.) of many different languages used by both genders.

(ii) $L_F$: total number of samples in each frame under consideration.
In this work, different values of $L_F$ (such as $L_F = 8, 16, 32, 64, 128$) were selected to investigate the effect of the frame length to the quality of the reconstructed speech by means of the absolute category rating-mean opinion score (ACR-MOS) and the segmental signal-to-noise ratio (SNRseg). Details of this effort are given in the subsequent section.

#### Computational steps

*Step 1.* Compute the total number of frames $N_F = N/L_F$.

*Step 2.* Divide the speech piece $X$ into frames $X_i$. In this case, the original speech is represented by the main frame vector $M_F^T = \lfloor X_1^T\ \ X_2^T\ \ \cdots\ \ X_{N_F}^T \rfloor$ of (5).

*Step 3.* For each frame $X_i$, compute the correlation matrix $R_i$.

*Step 4.* For each $R_i$, compute the eigenvalues $\lambda_{ik}$ in descending order with the corresponding eigenvectors.

*Step 5a.* Store the eigenvector which is associated with the maximum eigenvalue $\lambda_{ir} = \max\{\lambda_{i1}, \lambda_{i2}, \lambda_{i3}, \dots, \lambda_{iL_F}\}$ and simply refer to this signature vector with the frame index, as $S_{i1}$.
*Step 5b.* Compute the gain factor $C_{i1}$ in the LMS sense to approximate $X_i \approx C_{i1} S_{i1}$.

*Step 6.* Repeat Step 5 for all the frames ($i = 1, 2, \dots, N_F$). At the end of this loop, eigenvectors, which have maximum energy for each frame, will be collected.

*Step 7.* Compare all the collected eigenvectors obtained in Step 6 with an efficient algorithm. In this regard, Pearson correlation formula may be employed as described in Section 2.4. Then, eliminate the ones which exhibit similar patterns. Thus, generate the predefined signature set PSS $= \{S_{n_s}(n);\ n_s = 1, 2, \dots, N_S\}$ with reduced number of eigenvectors $S_{i1}$. Here, $N_S$ designates the total number of one of a kind signature patterns after the elimination. Remark: the above steps can be repeated for many different speech pieces to augment PSS.

*Step 8.* Compute the diagonal envelope matrix ($E_i$) for each $C_{i1} S_{i1}$ such that $e_{ir} = x_{ir}/(C_{i1} s_{i1r})$; $r = 1, 2, \dots, L_F$.

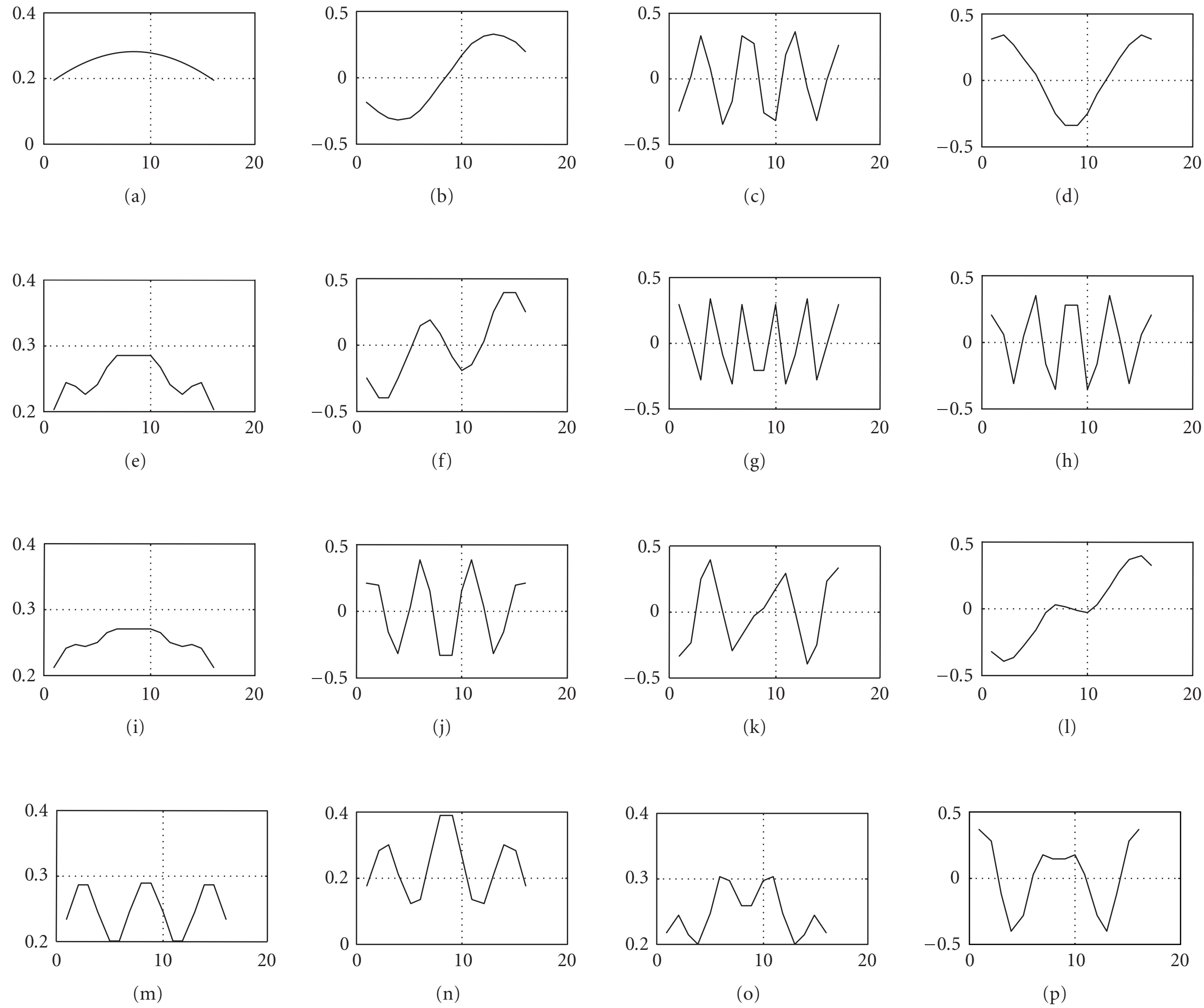


FIGURE 5: Unique patterns of some selected signature sequences ($L_F = 16$).

*Step 9.* Eliminate the envelope sequences which exhibit similar patterns with an efficient algorithm as in Step 7, and construct the predefined envelope set PES $= \{E_{n_e}(n);\ n_e = 1, 2, \ldots, N_E\}$; Here, $N_E$ denotes the total number of one of a kind unique envelope patterns.

Once PSS and PES are generated, then any speech signal can be reconstructed frame by frame ($X_{Ai} = C_i E_K S_R$) as implied by the main statement. It can be clearly seen that in this approach, the frame $i$ is reconstructed with three major quantities, namely, the gain factor $C_i$, the index $R$ of the predefined signature vector $S_R$ pulled from PSS, and the index $K$ of the predefined envelope sequence $E_K$ pulled from PES. $S_R$ and $E_K$ are determined to minimize the LMS error which is described by means of the difference between the original frame piece $X_i$ and its model $X_{Ai} = C_i E_K S_R$. Details of the reconstruction process are given in the following algorithm.

### 3.2. Algorithm 2: reconstruction of speech signals

#### Inputs

(i) Speech signal $\{X(n), n = 1, 2, \ldots, N\}$ to be modeled.
(ii) $L_F$: number of samples in each frame.
(iii) $N_S$ and $N_E$; total number of the elements in PSS and in PES, respectively. These integers are determined by Step 7 and Step 9 of Algorithm 1, respectively.
(iv) The predefined signature set PSS $= \{S_R; R = 1, 2, \ldots, N_S\}$ created utilizing Algorithm 1.
(v) The predefined envelope set PES $= \{E_K; K = 1, 2, \ldots, N_E\}$ created utilizing Algorithm 1.

#### Computational steps

*Step 1.* Divide $X$ into frames $X_i$ of length $L_F$ as in Algorithm 1. In this case, the original speech is represented by the main frame vector $M_F^T = \lfloor X_1^T \ \ X_2^T \ \ \cdots \ \ X_{N_F}^T \rfloor$ of (5).

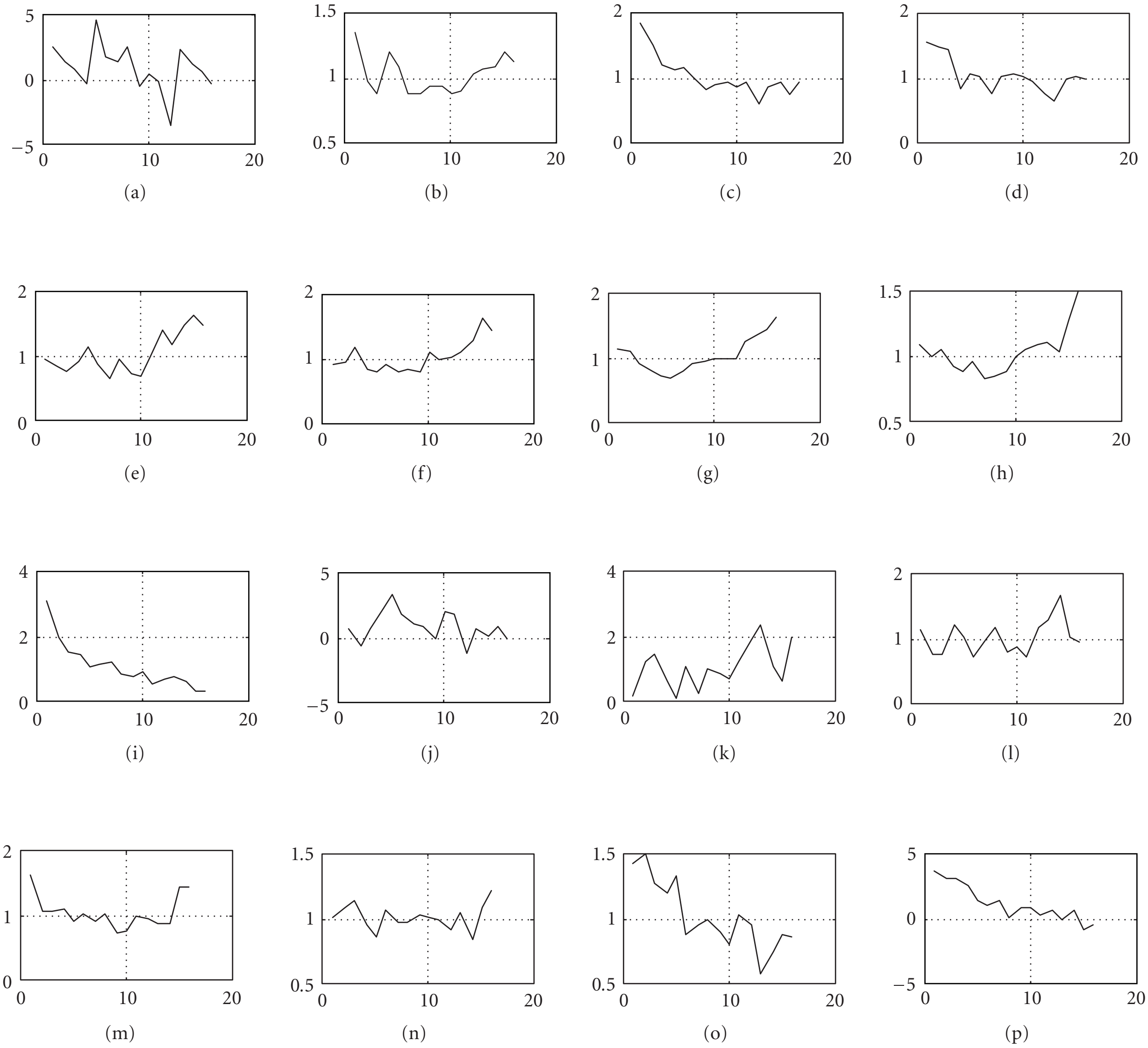

Figure 6: Unique patterns of some selected envelope sequences ($L_F = 16$).

*Step 2a.* For each frame $i$ pull an appropriate signature vector $S_R$ from PSS such that the distance or the total error $\delta_{\widetilde{R}} = \|X_i - C_{\widetilde{R}} S_{\widetilde{R}}\|^2$ is minimum for all $\widetilde{R} = 1, 2, \ldots, R, \ldots, N_S$. This step yields the index $R$ of the $S_R$. In this case, $\delta_R = \min\{\|X_i - C_{\widetilde{R}} S_{\widetilde{R}}\|^2\} = \|X_i - C_R S_R\|^2$.

*Step 2b.* Store the index number $R$ that refers to $S_R$, in this case, $X_i \approx C_R S_R$.

*Step 3a.* Pull an appropriate envelope sequence (or diagonal envelope matrix) $E_K$ from PES such that the error is further minimized for all $\widetilde{K} = 1, 2, \ldots, K, \ldots, N_E$. Thus, $\delta_K = \min\{\|X_i - C_R E_{\widetilde{K}} S_R\|^2\} = \|X_i - C_R E_K S_R\|^2$. This step yields the index $K$ of the $E_K$.

*Step 3b.* Store the index number $K$ that refers to $E_K$. It should be noted that at the end of this step, the best signature vector $S_R$ and the best envelope sequence $E_K$ are found by appropriate selections. Hence, the frame $X_i$ is best described in terms of the patterns of $E_K$ and $S_R$. That is, $X_i \approx C_R E_K S_R$.

*Step 4.* Having fixed $E_K$ and $S_R$, one can replace $C_R$ by computing a new gain factor $C_i = (E_K S_R)^T X_i / (E_K S_R)^T (E_K S_R)$ to further minimize the distance between the vectors $X_i$ and $C_R E_K S_R$ in the LMS sense. In this case, the global minimum of the error is obtained and it is given by $\delta_{\text{Global}} = \|X_i - C_i E_K S_R\|^2$. At this step, the frame sequence is approximated by $X_{Ai} = C_i E_K S_R$.

*Step 5.* Repeat the above steps for each frame to reconstruct speech as $M_{AF}^T = \lfloor X_{A1}^T \; X_{A2}^T \; \ldots \; X_{AN_F}^T \rfloor \approx M_F^T$.

In the following section, the new method of speech modeling is implemented for the frame lengths $L_F = 16$ and 128 to exhibit the usage of Algorithms 1 and 2 and the resulting speech quality are compared with the results of commercially available speech coding techniques G.726, LPC-10E, and also with our previous work [7].

## 4. INITIAL RESULTS ON THE IMPLEMENTATION OF THE NEW METHOD OF SPEECH REPRESENTATION

In this section, the speech reconstruction quality of the new method is compared with those of G.726 at 16 kbps and LPC-10E at 2.4 kbps providing (1 to 4) and (1 to 26.67) compression ratio, respectively. In this regard, the compression ratio (CR) is defined as $\mathrm{CR} = b_{\mathrm{org}}/b_{\mathrm{rec}}$; where $b_{\mathrm{org}}$ designates the total number of bits in representing the original signal and $b_{\mathrm{rec}}$ is the total number of bits which refers to the compressed version of the original. Finally, SYMPES is compared with the speech modeling technique presented in [7].

### 4.1. Comparison with G.726 (ADPCM) at 16 kbps

In order to make a fair comparison between G.726 at 16 kbps and the newly proposed technique, the input parameters of Algorithm 1 are arranged in such a way that Algorithm 2 of the reconstruction process yields $\mathrm{CR} = 4$. In this case, one only needs to measure the speech quality of the reconstructed signals as described below. In this regard, the speech pieces, which were given by the IPA Handbook and sampled with 8 KHz sampling rate were utilized to generate PSS and PES with $L_F = 16$ samples. In the generation process, all the available characteristic sentences (total of 253) from five different languages (English, French, German, Japanese, and Turkish) were employed. These sentences include consonants, conventions, introduction, pitch-accent, stress and accent, vowels (nasalized and oral), and vowel-length. Details are given in Table 1.

In this case, employing Algorithm 1, PSS was constructed with $N_S = 2048$ unique signature patterns. Similarly, PES was generated with $N_E = 57422$ unique envelopes. As described in Section 2.4 and step 7 of Algorithm 1, Pearson's similarity measure of (16) with $0.9 \leq \rho_{YZ} \leq 1$ was used in the elimination process. As a result of the above computations, $N_S$ and $N_E$ are represented with 11 and 16 bits, respectively. It was experienced that 5 bits were good enough to code the $C_i$. In conclusion, one ends up with a total number of $N_{BF} = 5 + 11 + 16 = 32$ bits to reconstruct the speech signals for each frame employing the newly proposed method. On the other hand, the original signal, coded with standard PCM (8 bits, 8 KHz sampling rate) is represented by $N_{B(PCM)} = 8 \times 16 = 128$ bits. Hence, both G.726 at 16 kbps and the new method provide $\mathrm{CR} = 4$ as desired. Under the given conditions, it is meaningful to compare the average ACR-MOS and the SNRseg, obtained for both G.726 and the new method. In the following section, ACR-MOS and SNRseg test results are presented.

It should be remarked that ideally one would expect to construct the universal predefined signature and envelope sets which are capable of producing all the existing sounds of languages. In this case, one may question the speech reproduction capability of PSS and PES derived using 253 different sound phrases mentioned above. Actually, we tried to enhance PSS and PES employing the other languages available in IPA. However, under the same elimination process implemented in Algorithm 1, we were not able to further increase the number of signature and the envelope patterns. Therefore, 253 sound phrases are good enough for the speech reproduction process of SYMPES. As a matter of fact, as it is shown by the following examples, the hearing quality of the new method (MOS $\approx$ 4.1) is much better than G.726 MOS $\leq$ 3.5). Hence, we confidently state that PSS and PES obtained for $L_F = 16$ provide good quality of speech reproduction.

#### 4.1.1. MOS and SNR assessment results: new method SYMPES versus G.726

In this section, mean opinion score and segmental signal-to-noise ratio results of SYMPES are presented and they are compared with those of G.726.

Mean opinion score tests: once PSS and PES are generated, the subjective test process contains three stages; collection of original speech samples, speech modeling or reconstruction, and the hearing quality evaluation of the reconstructed speech.

The original speech samples were collected from OGI, TIMIT, and IPA corpus databases [18, 21–23]. In this regard, we had the freedom to work with five languages namely; English, French, German, Japanese, and Turkish. Furthermore, for each language, we picked 24 different sentences or phrases which were uttered by 12 male and 12 female speakers. At this point, it is important to mention that PSS and PES should be universal (speaker and language independent) for any sound to be synthesized. Therefore, for the sake of fairness, we were careful not to use the same speech samples which were utilized in the construction PSS and PES. In the second stage of the tests, one has to model the selected speech samples using Algorithm 2. In the last stage, reconstructed speech pieces for both the new method and G.726 are evaluated by means of the subjective (ACR-MOS) and the objective (SNRseg) speech quality assessment techniques [24, 25].

Specifically, for subjective evaluation, we implemented the absolute category rating—mean opinion score (ACR-MOS) test procedure. In this process, firstly, the reconstructed speech pieces and then the originals are listened by several untrained listeners. Then, these listeners are asked to rate the overall quality of the reconstructed speech using five categories (5.0: excellent, 4.0: good, 3.0: fair, 2.0: poor, 1.0: bad). Eventually, one takes the average of the opinion scores of the listeners for the speech sample under consideration. An advantage of the ACR-MOS test is that subjects are free to assign their own perceptual impression to the speech quality. However, these freedom posses numerous disadvantages since the individual subject's goodness scales vary greatly. This variation can be a biased judgment. This bias could be avoided by using a large number of subjects. Therefore, as recommended by [26–29], we employed 40 (20 male and 20 female) subjects to come up with reliable ACR-MOS values.

TABLE 1: Language-based speech property distribution of the complete sample set provided by IPA utilized to form PSS and PES for $L_F = 16$.

| | | Languages | | | | |
|---|---|---|---|---|---|---|
| | | English | French | German | Japanese | Turkish |
| Speaker gender | | Female | Female | Male | Male | Male |
| Consonants | | 25 | 21 | 25 | 20 | 22 |
| Conventions | | 17 | — | 18 | 21 | 4 |
| Introduction | | — | — | 4 | — | — |
| Pitch-accent | | — | — | — | 6 | — |
| Stress-and-accent | | — | — | 1 | — | 3 |
| Vowels | Nasalized | 15 | 3 | 19 | 5 | 8 |
| | Oral | | 12 | | | |
| Vowel-length | | — | — | — | 4 | — |
| Subtotal number of words | | 57 | 36 | 67 | 56 | 37 |
| Total number of words | | | | 253 | | |

In order to assess the objective quality of the reconstructed speech signals, the SNRseg is utilized. Here, in this work, each segment is described over 10 frames of length $L_F = 16$ or equivalently each segment consists of $K_F = 160$ samples. Then, SNRseg is given by

$$\mathrm{SNR}_{\mathrm{seg}} = \frac{1}{T_F} \sum_{j=0}^{T_F-1} 10 \log_{10} \left[ \frac{\sum_{n=m_j-K_F+1}^{m_j} [x(n)]^2}{\sum_{n=m_j-K_F+1}^{m_j} [x(n) - \hat{x}(n)]^2} \right]. \tag{17}$$

Let $N$ be the total number of samples in the speech piece to be reconstructed. Then, in (17) $T_F = N/K_F$; $j$ designates the frame index; $n$ is the sample number in frame $j$; $m_0 = K_F$; $m_j = \mathrm{j}\mathrm{K}_F$. It should be noted that the indices $m_0, m_1, \ldots, m_{T_F-1}$ refer to the "end points" of each segment placed in the speech piece to be reconstructed.

The ACR-MOS test results and computed values of SNRseg for the reconstructed speech pieces are summarized in Table 2.

If we compute the average ACR-MOS and SNRseg values over the languages, one can clearly see that the new method provides much better speech quality over G.726. In this case, we can say that the proposed method yields almost toll quality (MOS $\approx 4.1$) whereas G.726 is considered to yield communication quality (MOS $\approx 3.5$). To provide visual comprehension, the original and the reconstructed waveforms of the five speech waveforms corresponding to five different sentences in five languages uttered by male speakers are depicted in Figure 7. Similarly, in Figure 8, speech waveforms uttered by female speakers are shown.

As it can be deduced from Figure 7, the visual difference between the original and the reconstructed waveforms are negligible, which verifies the superior results presented in Table 2 for the newly proposed speech modeling technique. This completes the comparison at the low compression rate (CR = 4).

It should be mentioned that similar comparisons were also made with G.726 at 24, 32, and 48 kbps. For these cases proposed method yields slightly better results over G.726. For example, the new method with $L_F = 8$ corresponds to G.726 at 32 kbps. In this case, while G.726 results in $\mathrm{SNR}_{G.726-32} \approx 25$ dB, the new method gives SNR $\approx$ 26 dB. Since the difference is negligible, details are omitted here.

Let us now comment on the noise robustness of SYMPES.

#### *4.1.2. Comments on the noise robustness of SYMPES*

SYMPES directly builds a mathematical model for the speech signal regardless it is noisy or not. Therefore, one expects to end up with a similar noise level in the reconstructed speech as in the original. In fact, a subjective noise test was run to observe the effect of the noisy environment to the robustness of SYMPES. In this regard, a noise free speech piece was mixed with 1.2 dB white noise; then it was reconstructed using SYMPES of $L_F = 16$. The test was run among 5 male and 5 female untrained listeners. They were asked to rate the noise level of the reconstructed speech relative to the original, under three categories namely "no change in the noise level," "reduced noise level," and "increased noise level". Seven of the listeners confirmed that the noise level of the reconstructed speech was not changed. Two of the female subjects said that the noise level was slightly reduced, and one of the male listener asserted that noise level was slightly increased. In this case, we can safely state that "SYMPES is not susceptible to the noise level of the environment." Furthermore, any noise level which is built on the original signal can be reduced by post-filtering the reconstructed signal. As a matter of fact it was experienced that both the background noise due to reconstruction process and the environmental noise were reduced significantly by using a moving average post-filter.

At this point, it may be meaningful to make a further comparison at high compression rates such as CR = 25 or higher. For this purpose, voice excited LPC-10E which yields CR = 26.67 may be considered as outlined in the following section.

Table 2: Subjective and objective speech quality scores for G726 and the new method.

| Language | Speaker gender | Number of speech pieces | Bit rate [kbps] | ACR-MOS (G.726) ADPCM | ACR-MOS SYMPES | SNRseg [dB] (G.726) ADPCM | SNRseg [dB] SYMPES |
|---|---|---|---|---|---|---|---|
| English | Male | 12 | 16 | 3.417 | 4.124 | 7.4014 | 12.4033 |
| | Female | 12 | | 3.419 | 4.109 | 7.4289 | 12.1969 |
| French | Male | 12 | 16 | 3.413 | 4.111 | 7.3513 | 12.2083 |
| | Female | 12 | | 3.422 | 4.099 | 7.4396 | 12.0518 |
| German | Male | 12 | 16 | 3.386 | 4.051 | 6.9072 | 11.4075 |
| | Female | 12 | | 3.371 | 4.036 | 6.6886 | 11.2053 |
| Japanese | Male | 12 | 16 | 3.422 | 4.167 | 7.4599 | 12.9719 |
| | Female | 12 | | 3.668 | 4.272 | 11.1795 | 14.4533 |
| Turkish | Male | 12 | 16 | 3.453 | 4.040 | 7.9029 | 11.2603 |
| | Female | 12 | | 3.433 | 4.010 | 7.6134 | 10.8320 |
| Average scores | | | | 3.440 | **4.102** | 8.000 | **12.000** |

### 4.2. Comparison with voice excited LPC-10E (2.4 kbps)

Standard voice excited LPC-10E employs 20 msec speech frames coded with 48 bits which corresponds to 2.4 kbps. On the other hand, using standard PCM, these time frames contain 160 samples represented by 1280 bits. Thus, the compression rate of LPC-10E is $CR_{LPC} = 1280/48. = 26.67$. In order to make a fair comparison, parameters of the new method have to match to that of LPC-10E. First of all, PSS and PES must be regenerated accordingly. In this regard, we can say that one needs to deal with a multitudinous variety of many "signature and envelope" sets to enhance the language & speaker independency for the long speech frame lengths such as $L_F = 128$. However, it should be recalled that this was not the case for $L_F = 16$. So, as described in Section 4.1, we utilized the rich speech samples collection of IPA [18] with 890 different characteristic sentences in 17 different languages (English, French, German, Japanese, Turkish, Amharic, Arabic, Irish, Sindhi, Cantonese, Czech, Bulgarian, Dutch, Hebrew, Catalan, Galician, and Croatian) (see Table 3). Choosing $L_F = 128$ and $0.9 \le \rho_{YZ} \le 1$, Algorithm 1 returns with $N_S = 32768$ signature and $N_E = 131072$ envelope patterns of one kind. Clearly, it is sufficient to represent $N_S$ and $N_E$ with 15 and 17 bits, respectively. As was the case before, the gain factor $C_i$ is also represented with 5 bits. In this case, each frame of 128 samples is represented by total number of $N_{BF} = 5+15+17 = 37$ bits. Thus, the compression ratio of the new method becomes $CR = 128 \times 8/37 = 27.68$ which is even higher than $CR_{LPC} = 26.67$. In the following section it is shown that the new method yields superior speech quality over voice excited LPC-10E.

#### 4.2.1. MOS test results: SYMPES versus voice excited LPC-10E

As described in Section 4.1.1, after the formation of PSS and PES with $L_F = 128$ samples, we run the ACR-MOS test with the same speech set given by Table 2. The test results are summarized in Table 4.

A close examination of Table 4 reveals that SYMPES results in superior speech quality over voice excited LPC-10E for all the languages under consideration.

Just for the sake of visual inspection an original and a reconstructed speech signals are depicted in Figure 9 for comparison. A close examination of Figure 9 validates the superior reconstruction ability of SYMPES over voice excited LPC-10E.

#### 4.2.2. Comparison of SYMPES with CS-ACELP

It is important to mention that one may conceptually link SYMPES with the other code excited linear predictive (CELP) methods such as conjugate structure-algebraic CELP (CS-ACELP) at 8 kbps (or G.729 at 8 kbps).

CS-ACELP utilizes two stage LBG vector quantization with fixed[2] and adaptive[3] codebooks [30]. In this regard, each speech frame of 10 msec is described in terms of the indices of the fixed and adaptive codes and the gain factor and they are represented with a total of 80 bits which corresponds to a compression ratio of $CR_{CS\text{-}ACELP} = 8$. This process may resemble the procedure described by SYMPES. Fixed and adaptive codes of CS-ACELP may be related to the signature and the envelope sequences of SYMPES respectively; but it should be kept in mind that SYMPES does not include any adaptive quantity beyond the gain factor. Furthermore, CS-ACELP is an LPC technique which takes the error or the residual into account in an additive manner whereas SMYPES literally produces a simple but a nonlinear frame model by multiplying three major quantities so that $X_{Ai} = f(C_i, E_K, S_R) = C_i E_K S_R$. In this representation, the envelope matrix $E_K$ works on the signature vector $S_R$ as a multiplier to reduce the modeling error in a nonlinear manner. Clearly, it is not possible to find a one-to-one correspondence between the SYMPES and the CS-ACELP,

[2] Voice excitations.
[3] Line spectral pairs (LSP) envelope parameters.

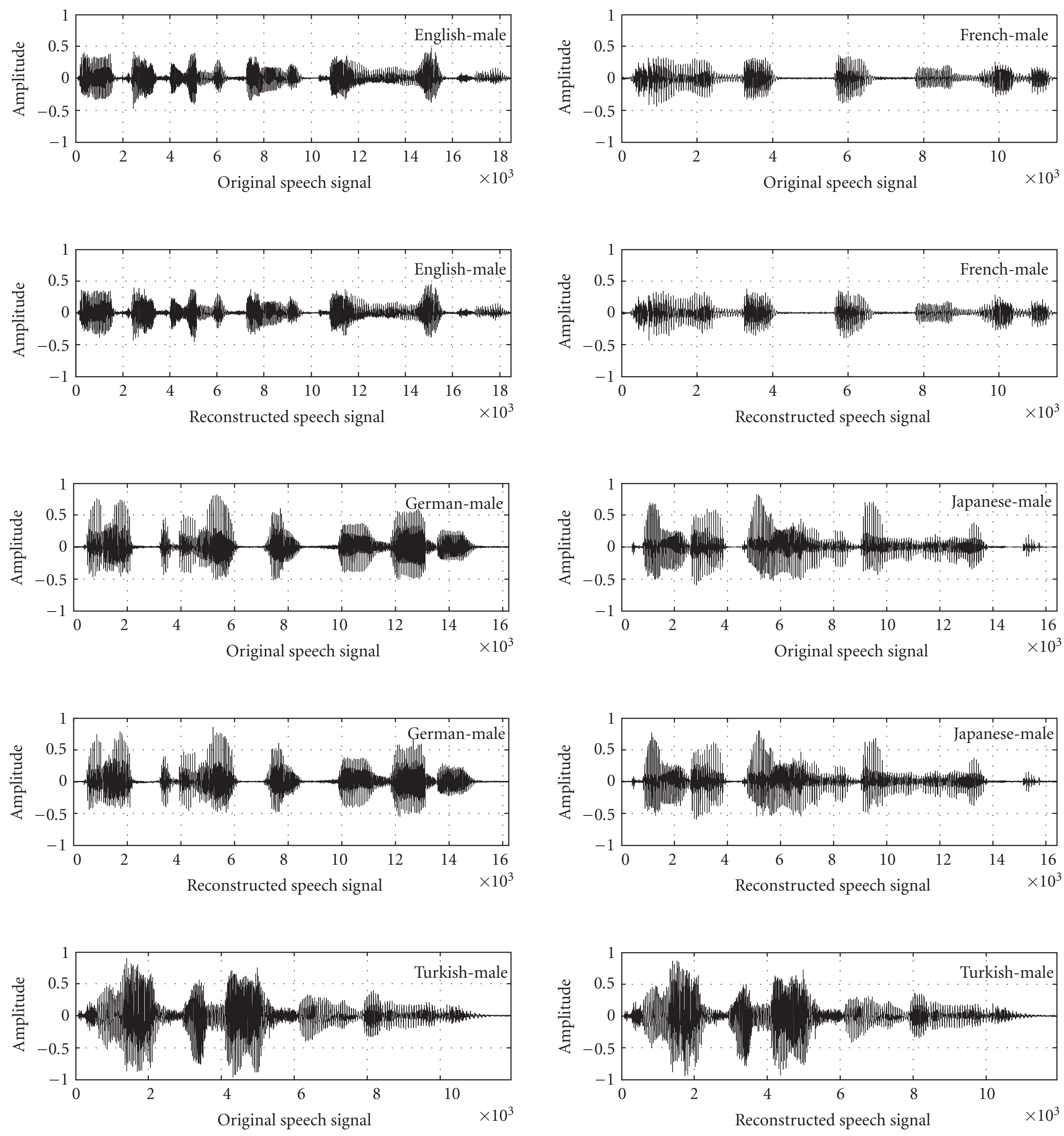


Figure 7: Original and reconstructed speech waveforms using the new method for English, French, German, Japanese, and Turkish sentences uttered by male speakers.

since they differ in nature with respect to both model[4] and domain[5]. On the other hand, the gain factor $C_i$ of SYMPES plays the same role as in CS-ACELP to further reduce the error between the original and the approximated speech frames in the LMS sense. Similar MOS tests of Section 4.2.1 were also run to compare SYMPES at $L_F = 32$[6] with CS-ACELP at 8 kbps. It was found that SYMPES yields the

[4] Linear model of CS-ACELP versus nonlinear model of SYMPES.

[5] Transform domain of CS-ACELP versus discrete time domain of SYMPES.

[6] SYMPES $L_F = 32$ with 8 KHz sampling rate yields the compression ration of CR = 8 as in CS-ACELP at 8 kbps.

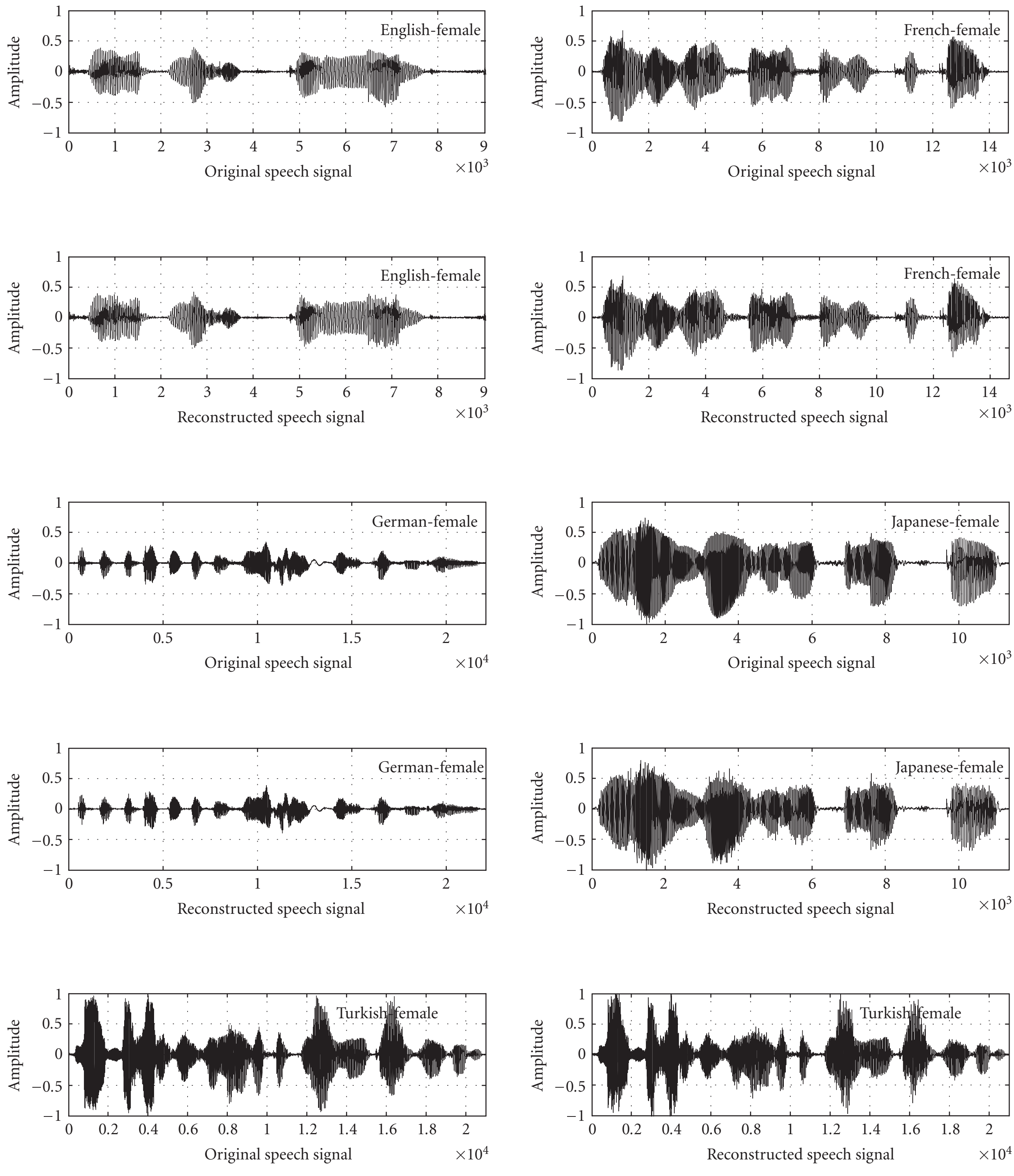


Figure 8: Original and reconstructed speech waveforms using the new method for English, French, German, Japanese, and Turkish sentences uttered by female speakers.

average $\text{MOS}_{\text{SYMPES}} = 3.72$ in contrast with CS-ACELP giving the average $\text{MOS}_{\text{CS-ACELP}} = 3.70$. Details are omitted here since the hearing quality difference between the two methods is negligible.

Based on the experimental results of this research, we conclude that SYMPES provides much better hearing quality than that of commercially available G.726 and CELP coding techniques at high compression rates ($\text{CR} \gg 8$). At low

Table 3: Language-based speech property distribution of the complete sample set provided by IPA utilized to form PSS and PES for $L_F = 128$.

| Language | Speaker gender | Consonant | Convention | Vowels | Stress and accent | Introduction | Pitch-accent | Vowel-length | Assimilation | Geminatives |
|---|---|---|---|---|---|---|---|---|---|---|
| English | Female | 25 | 17 | 15 | — | — | — | — | — | — |
| French | Female | 21 | — | Nasalized 3<br>Oral 12 | — | — | — | — | — | — |
| German | Male | 25 | 18 | 19 | 1 | 4 | — | — | — | — |
| Japanese | Male | 20 | 21 | 5 | — | — | 6 | 4 | — | — |
| Turkish | Male | 22 | 4 | 8 | 3 | — | — | — | — | — |
| Amharic | Male | 35 | — | 11 | — | — | — | — | — | — |
| Arabic | Male | 29 | — | 8 | — | — | — | — | — | — |
| Irish | Female | 44 | — | 14 | — | — | — | — | — | — |
| Sindhi | Male | 46 | — | 10 | — | — | — | — | — | — |
| Cantonese | Male | 19 | — | Diphthongs 11<br>Monophthongs 32 | — | — | — | — | — | 9 |
| Czech | Female | 25 | — | 13 | | 5 | — | — | 3 | — |
| Bulgarian | Female | 22 | — | 8 | 2 | — | — | — | — | — |
| Dutch | Female | 23 | — | 22 | 4 | — | — | — | — | — |
| Hebrew | Male | 22 | — | 5 | 2 | — | — | — | — | — |
| Catalan | Male | 23 | 21 | Diphthongs 8<br>Stressed 7<br>Unstressed 3 | 7 | — | — | — | — | — |
| Galician | Male | 21 | 22 | 7 | 23 | — | — | — | — | — |
| Croatian | Female | 25 | 10 | 1<br>Long 7<br>Short 5 | 20 | 3 | — | — | — | — |
| Subtotal number of words | | 447 | 113 | 234 | 62 | 12 | 6 | 4 | 3 | 9 |
| Total number of words | 890 | | | | | | | | | |

Table 4: Subjective speech quality scores for LPC-10E and the new method.

| Language | Speaker gender | Number of speech pieces | ACR-MOS LPC-10E 2.4 kbps | ACR-MOS SYMPES 2.3125 kbps |
|---|---|---|---|---|
| English | Male | 12 | 2.490 | 3.384 |
| | Female | 12 | 2.395 | 3.455 |
| French | Male | 12 | 2.520 | 3.374 |
| | Female | 12 | 2.409 | 3.435 |
| German | Male | 12 | 2.540 | 3.363 |
| | Female | 12 | 2.410 | 3.411 |
| Japanese | Male | 12 | 2.460 | 3.359 |
| | Female | 12 | 2.427 | 3.603 |
| Turkish | Male | 12 | 2.610 | 3.396 |
| | Female | 12 | 2.452 | 3.418 |
| Average scores | | | **2.471** | **3.420** |

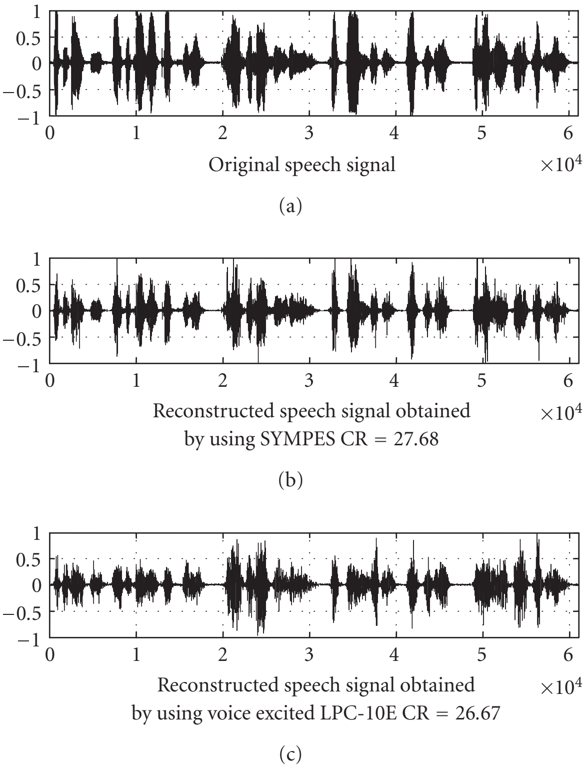


Figure 9: Original and the reconstructed speech signals for visual inspection and comparison of the new method of speech modeling SYMPES with LPC-10E.

compression rates (CR $\leq$ 8) however, SYMPES yields either slightly better or almost the same speech quality like the others.

### 4.3. Comparison of SYMPES with our previous results given by [7]

First of all in [7], the results were given on the predefined signature set which was generated based on selected 500 words from Turkish Language, which in turn makes the speech model very restricted; whereas in this work, complete speech pieces of OGI, TIMIT, and IPA Handbook were utilized to generate predefined signature and envelope sets which are supposed to yield rather universal results and make SYMPES speaker and language independent.

Moreover, in [7], envelope sequences which improve the hearing quality tremendously were not used at all. Hence, here in this work, results of [7] were pretty much generalized and hearing quality of the reconstructed speech signals is significantly enhanced. As a matter of fact, no matter what the frame length and the compression ratio is, in the reconstruction process, mean opinion scores presented in [7] were below 2.8 out of 5, whereas in this work, in all the examples, they are well above 3.4. Therefore, we can simply state that SYMPES is the generalized and the improved version of the speech model method presented in [7].

## 5. CONCLUSIONS

In this paper, a novel systematic procedure referred to as "SYMPES" is presented to model speech signals frame by frame by means of the so-called predefined "signature and envelope" patterns. In this procedure, the reconstructed speech frame $X_{Ai}$ is described by multiplying three major quantities, namely, the gain factor $C_i$, the frame signature vector $S_R$, and the diagonal envelope matrix $E_K$ or in short as $X_{Ai} = C_i E_K S_R$. Signature and envelope patterns are selected from the corresponding PSS and PES that are formed through the use of a variety of speech samples included in the IPA Handbook. These sets are almost universal. That is to say, they are speaker and language independent. In the synthesis process, each speech frame is fully identified with the gain factor $C_i$ and the indices $R$ and $K$ of the predefined signature and the envelope patterns, respectively.

The subjective and objective test assessments reveal that the hearing quality of SYMPES is slightly better at low compression rates (CR $\leq$ 8) than that of G.726 (16, 24, 32, and 48 kbps) and CS-ACELP (8 kbps). At higher compression rates (CR $\gg$ 8), SYMPES results in superior hearing quality over G.726 and LPC techniques. One should note that this high rate of compression is purchased at the expense of the computational efforts to determine the gain factors as well as to identify the proper signature and envelope patterns in the search process. In this regard, computational lag may be disregarded by an appropriate buffering operation.

As far as digital communication systems are concerned, SYMPES may be considered as a coding scheme. In this case, once the PSS and PES are created and stored, one only needs to transmit the $C_i$ with the relevant indices $R$ and $K$. For example, if SYMPES with $L_F = 128$ is used, then a substantial saving in the transmission-bandwidth (CR = 27.68) with good quality of speech is achieved.

It is interesting to note that the new method of speech modeling presented in this paper may be employed for speech recognition purposes as described in [31]. It may be used to model biomedical signals such as electrocardiograms and electromyograms as well. Initial results of these works are given in [32, 33]. In future research, we hope to improve the results of [31–33] and the computational efficiency of SYMPES.

## ACKNOWLEDGMENT

This work is sponsored by the research unit of Istanbul University, Istanbul, Turkey under the Contracts No. UDP-440/10032005 and 400/03062005.

## REFERENCES

[1] A. S. Spanias, "Speech coding: a tutorial review," *Proceedings of the IEEE*, vol. 82, no. 10, pp. 1541–1582, 1994.

[2] S. Watanabe, "Karhunen-Loeve expansion and factor analysis; theoretical remarks and applications," in *Transactions of the 4th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pp. 635–660, Czechoslovak Academy of Sciences, Prague, Czech Republic, 1965.

[3] G. Varile and A. Zampolli, *Survey of the State of the Art in Human Language Technology*, chapter 10.2: Transmission and Storage (B. S. Atal and N. S. Jayant), Cambridge University Press, Cambridge, UK, 1998.
[4] A. M. Karaş and B. S. Yarman, "A new approach for representing discrete signal waveforms via private signature base sequences," in *Proceedings of the IEEE European Conference on Circuit Theory and Design*, pp. 875–878, Istanbul, Turkey, August 1995.
[5] A. M. Karaş, *Characterization of electrical signals by using signature base functions*, Ph.D. thesis, Department of Electrical and Computer Engineering, Institute of Science, Istanbul University, Istanbul, Turkey, January 1997, Advisor: Professor B. S. Yarman.
[6] R. Akdeniz and B. S. Yarman, "Turkish speech coding by signature base sequences," in *Proceedings of the International Conference on Signal Processing Applications & Technology (ICSPAT '98)*, pp. 1291–1294, Toronto, Canada, September 1998.
[7] R. Akdeniz and B. S. Yarman, "A novel method to represent speech signals," *Signal Processing*, vol. 85, no. 1, pp. 37–50, 2005.
[8] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–498, 1933.
[9] E. Oja, "A simplified neuron model as a principal component analyzer," *Journal of Mathematical Biology*, vol. 15, no. 3, pp. 267–273, 1982.
[10] I. T. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics, Springer, New York, NY, USA, 1933.
[11] A. N. Akansu and R. A. Haddad, *Multiresolution Signal Decomposition*, Academic Press, San Diego, Calif, USA, 1992.
[12] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, London, UK, 1990.
[13] A. J. Newman, "Model reduction via the Karhunen Loeve expansion part I: an exposition," Tech. Rep. ISR T.R.96-32, Institute of Systems Research, College Park, Md, USA, April 1996.
[14] G. Strang, *Linear Algebra and Its Applications*, Academic Press, New York, NY, USA, 1980.
[15] Ü. Güz, *A new approach in the determination of optimum signature base functions for Turkish speech*, Ph.D. thesis, Department of Electrical and Computer Engineering, Institute of Science, Istanbul University, Istanbul, Turkey, 2002, Advisor: Professor B. S. Yarman.
[16] Ü. Güz, B. S. Yarman, and H. Gürkan, "A new method to represent speech signals via predefined functional bases," in *Proceedings of the IEEE European Conference on Circuit Theory and Design*, vol. 2, pp. 5–8, Espoo, Finland, August 2001.
[17] Ü. Güz, H. Gürkan, and B. S. Yarman, "A novel method to represent the speech signals by using language and speaker independent predefined functions sets," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, vol. 3, pp. 457–460, Vancouver, BC, Canada, May 2004.
[18] IPA, *Handbook of the International Phonetics Association: A Guide to the Use of the International Phonetic Alphabet*, Cambridge University Press, Cambridge, UK, 1999.
[19] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, no. 11, pp. 559–572, 1901.
[20] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, 1980.
[21] OGI Multi-Language Telephone Speech Corpus, CD-ROM, Linguistic Data Consortium.
[22] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*, Prentice Hall, Englewood Cliffs, NJ, USA, 1988.
[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic speech corpus," Tech. Rep. NISTIR 4930, U.S. Department of Commerce, NIST, Computer Systems Laboratory, Washington, DC, USA, 1993.
[24] ITU-T Recommendation G.726; 40, 32, 24, 16 kbit/s ADPCM, Geneva, (12/90).
[25] ITU-T Appendix III to ITU-T Recommendation G.726; General aspects of digital transmission systems-comparison of ADPCM algorithms, Geneva, (05/94).
[26] ITU-T Recommendation P.861; Series P: Telephone transmission quality methods for objective and subjective assessment of quality-objective quality measurement of telephone band (300-3400 Hz) speech codecs, Geneva, (08/96).
[27] ITU-T Recommendation P.830; Telephone transmission quality methods for objective and subjective assessment of quality-subjective performance assessment of telephone-band and wideband digital codecs, Geneva, (02/96).
[28] W. D. Voiers, "Methods of predicting user acceptance of voice communication systems," Final Report DCA100-74-C-0056, July 1976.
[29] ITU-T Recommendation P.800; Series P: Telephone transmission quality methods for objective and subjective assessment of quality-methods for subjective determination of transmission quality, Geneva, (08/96).
[30] ITU-T Recommendation G.729; Coding of speech at 8 kbit/s using CS-ACELP.
[31] Ü. Güz, H. Gürkan, and B. S. Yarman, "A new speech signal modeling and word recognition method by using signature and envelope feature spaces," in *Proceedings of the IEEE European Conference on Circuit Theory and Design*, vol. 3, pp. 161–164, Cracow, Poland, September 2003.
[32] B. S. Yarman, H. Gürkan, Ü. Güz, and B. Aygün, "A new modeling method of the ECG signals based on the use of an optimized predefined functional database," *Acta Cardiologica - An International Journal of Cardiology*, vol. 58, no. 3, pp. 59–61, 2003.
[33] H. Gürkan, Ü. Güz, and B. S. Yarman, "A novel representation method for electromyogram (EMG) signal with predefined signature and envelope functional bank," in *Proceedings of the IEEE International Symposium on Circuits and Systems*, vol. 4, pp. 69–72, Vancouver, BC, Canada, May 2004.

**Ümit Güz** graduated from Istanbul Pertevniyal High School in 1988 and Department of Computer Programming, Yıldız Technical University, Istanbul, Turkey in 1990. He received the B.S. degree with high honors from the Department of Electronics Engineering, College of Engineering, Istanbul University, Istanbul, Turkey in 1994. He received M.S. and Ph.D. degrees in electronics engineering from the Institute of Science, Istanbul University, Istanbul, Turkey, in 1997 and 2002, respectively. From 1995 to 1998 he was a Research and Teaching Assistant in the Department of Electronics Engineering, Istanbul University. He has been an Instructor in the Department of Electronics Engineering, Engineering Faculty, Işık University, Istanbul, Turkey, since 1998. He is awarded with postdoctoral research fellowship by The Scientific and Technical Research Council of Turkey

(TÜBİTAK) in 2006. He is accepted as an International Fellow by the SRI (Stanford Research Institute)-International Speech Technology and Research (STAR) Laboratory in 2006. He is awarded with the J. William Fulbright Post-Doctoral Research Fellowship in 2007. He is accepted as an International Fellow by the International Computer Science Institute (ICSI) Speech Group at the University of California, Berkeley in 2007. His research interest covers speech modeling, speech coding, speech compression, automatic speech recognition, natural language processing, and biomedical signal processing.

**Hakan Gürkan** received the B.S., M.S., and Ph.D. degrees in electronics and communication engineering from the Istanbul Technical University, Istanbul, Turkey, in 1994, 1998, and 2005, respectively. He was a Research Assistant in the Department of Electronics Engineering, Engineering Faculty, Işık University, Istanbul, Turkey. He has been an instructor in the Department of Electronics Engineering, Engineering Faculty, Işık University, Istanbul, Turkey, since 2005. His current interests are in digital signal processing, mainly with biomedical and speech signals modeling, representation, and compression.

**Binboga Sıddık Yarman** received the B.S. degree in electrical engineering from Istanbul Technical University, Turkey (1974); M.E.E.E. degree from Electro-Math Stevens Institute of Technology Hoboken, NJ, 1977; Ph.D. degree in EE-Math from Cornell University, Ithaca, NY, 1981. He was a Member of the Technical Staff, Microwave Technology Centre, RCA David Sarnoff Research Center, Princeton, NJ (1982–1984); Professor, Alexander Von Humboldt Fellow, Ruhr University, Bochum, Germany (1987–1994); Founding Director, STFA Defense Electronic Corp., Turkey (1986–1996); Professor, Chair, Defense Electronics, Director, Technology and Science School, Istanbul University (1990–1996); Founding President of Işık University, Istanbul, Turkey (1996–2004); Chief Advisor to Prime Ministry Office, Turkey (1996–2000); Chairman of the Science Commission, Turkish Rail Roads, Ministry of Transportation (2004). He obtained the Young Turkish Scientist Award, National Research Council of Turkey (NRCT) (1986); and Technology Award of NRCT (1987); International Man of the Year in Science and Technology, Cambridge Biography Center of U.K. (1998). He was a Member of the Academy of Science of New York (1994), Fellow of IEEE. He is the author of more than 100 papers, 4 US patents. Fields of interests include design of matching networks and microwave amplifiers, mathematical models for speech and biomedical signals. He has been back to Istanbul University since October 2004 and spending his sabbatical year of 2006–2007 at Tokyo Institute of Technology, Tokyo, Japan.