A NEW CODING METHOD FOR SPEECH AND AUDIO SIGNALS

Ümit Güz[1], Hakan Gürkan[1], and B. Sıddık Yarman[2]

[1] Işık University, Engineering Faculty, Dept. of Electronics Engineering, Buyukdere Cad., 34398, Maslak, Istanbul, Turkey
[2] Istanbul University, Engineering Faculty, Dept. of Electrical and Electronics Engineering, Avcılar, Istanbul, Turkey
guz@isikun.edu.tr, hakan@isikun.edu.tr, yarman@istanbul.edu.tr

ABSTRACT

In this paper a new representation or modeling method of speech signals is introduced. The proposed method is based on the generation of the so-called Predefined Signature $S=\{S_R\}$ and Envelope vector $E=\{E_K\}$ Sets (PSEVS). These vector sets are speaker and language independent. In this method, once the speech signals are divided into frames with selected lengths, then each frame signal piece X_i is reconstructed by means of the mathematical form of $X_i=C_iE_KS_R$. In this representation, C_i is called the frame coefficient, S_R and E_K are the vectors properly assigned from the PSEVS respectively. It is shown that the proposed method provides fast reconstruction and substantial compression ratio with acceptable hearing quality.

1. INTRODUCTION

The primary objective of the speech coding is to represent the speech signal with a minimum number of bits while maintaining its perceptual quality [1].

Over the decades, a number of new algorithms based on the use of numerical, mathematical, statistical and heuristic methodologies have been proposed for representation, coding or compression of the speech signals. These algorithms also utilize time or frequency domain and spectral properties of the speech signals. The proposed algorithms use LPC and its derivative based techniques such as RELP, MLPC, CELP, VSELP, VSECLP etc., PCM based techniques such that DPCM, ADPCM etc., Transform Domain Techniques such as Wavelet Transform, DCT, Principal Component Analysis (PCA) or Karhunen Loeve Decomposition (KLD) etc., and hybrid techniques. Some of these methods are able to produce high quality speech at high enough bit rates. On the other hand, some of the other methods produce intelligible speech at much lower bit rates, but the level of the speech quality by means of its naturalness is also much lower [1].

In this context, the aim of the proposed method is to achieve speech coding at low bit rates without an objectionable loss of speech quality in terms of its intelligibility and naturalness and to bring new capabilities in recent coding and compression technology.

In our previous techniques given by [2-9], one would first examine the signal in terms of its physical features, and then find some specific waveforms to best describe the signals, which are called Signature Base Functions (SBF). The SBF of the speech signals are obtained by using energy compaction property of the PCA. The PCA also provides an optimal solution via minimization of the error in the Least Mean Square (LMS) sense. In the new method presented in this paper, the results of [2-9] have been significantly improved by introducing the concept of the Envelope Vector in the representation of speech signals. Thus the new mathematical form of the frame signal X_i is proposed as $X_i=C_iE_KS_R$ where C_i is a real constant called the frame coefficient, S_R and E_K are properly extracted from the so-called Predefined Signature Vector Set $S=\{S_R\}$ and Predefined Envelope Vector Set $E=\{E_K\}$ or in short PSVS and PEVS respectively [9]. Eventually, it has been exhibited that PSVS and PEVS are speaker and language independent. It should also be mentioned that if the proposed modeling technique is employed in communication, it results in substantial reduction in transmission bandwidth. If it is used for digital recording, it provides huge savings in the storage area.

2. DETAILS OF THE NEW METHOD

The speech signals are non-stationary and at best they can be considered as quasi-stationary over short segments. The statistical and spectral properties of the speech signals are thus defined over short segments [1]. In another words, 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 (original) speech signal with N samples. This signal is segmented into non-overlapping vectors that are called frames and the frame-length corresponds to the dimension of the frame vectors. Let X(n) be analyzed frame by frame and $X_i(n)$ represents the i^{th} frame of the original signal with length L_F. Then, we make the following Main-Statement which constitutes the basis of the proposed technique in this work.

2.1 Main Statement

In this section, the mathematical form to approximate the speech signals over a finite interval is given in discrete time domain. For any time frame i, the sampled speech signal which is given by the vector X_i with length L_F, can be approximated as

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

where C_i a real constant, $K\in\{1,2,\ldots,N_E\}$, $R\in\{1,2,\ldots,N_S\}$; K, R, N_E, and N_S and are integers. The vector $S_R{}^T=\left[s_{R1}, s_{R2}, s_{R3},\ldots, s_{RL_F}\right]$ is generated utilizing the statistical behaviour of the speech signals and it includes basic characteristics of the original frame under consideration in broad sense. Furthermore, it will be shown that the quantity C_iS_R carries almost maximum energy of X_i in the LMS sense. E_K is $(L_F\times L_F)$ diagonal matrix such that $E_K = diag\left\lfloor e_{K1} \quad e_{K2} \quad e_{K3} \quad \ldots \quad e_{KL_F}\right\rfloor$. E_K acts as an envelope term on the quantity C_iS_R which also reflects the statistical properties of the speech signal under consideration. The integer L_F designates the total number of samples in the frame i.

Generation of the vector S_R and the matrix E_K will be given in the verification of the main statement. Based on the main statement, we can make the following definitions.

Definition 1: The S_R (or equivalently the sequence $\{s_{Rj}$; $j=1,2,\ldots,L_F\}$) is called the Signature Vector (or Signature Sequence) since it reflects the basic statistical characteristics of the original signals and carries almost maximum energy of the original signal contained in the frame i within a constant C_i.

Definition 2: The diagonal matrix E_K (or equivalently the sequence $\{e_{Kj}$; $j=1,2,\ldots,L_F\}$) is called the Envelope Matrix (or equivalently envelope sequence) since it matches the original frame vector X_i to C_iS_R.

Definition 3: The real constant C_i is called the Frame Coefficient.

In order to verify the main statement; let us proceed as follows.

2.2 Verification of the Main Statement

A long-sampled speech signal sequence x(n) can be written as

$$x(n)=\sum_{i=1}^{N} x_i\delta_i(n-i) \tag{2a}$$

In this equation, $\delta_i(n)$ represents the unit sample; x_i designates the amplitude of the sequence x(n) of length N. For very long speech signals one may consider the limit case such that $N\rightarrow\infty$. x(n) can also be given employing the vector/matrix notation.

$$X^T=\left[x(1)\quad x(2)\quad \ldots\quad x(N)\right]=\left[x_1\quad x_2\quad \ldots\quad x_N\right] \tag{2b}$$

In this representation X is called the Main Frame Vector and it is divided into sub-frames with equal lengths, having, for example, 16, 32, 64, or 128 samples etc. The Main Frame Matrix that is represented by M_F is obtained by means of the sub-frame vectors.

$$M_F=\left[X_1\quad X_2\quad \ldots\quad X_{N_F}\right] \tag{3}$$

where

$$X_i^{\,T}=\left[x_{(i-1)L_F+1}\quad x_{(i-1)L_F+2}\quad \ldots\quad x_{iL_F}\right],\ i=1,2,3,\ldots,N_F \tag{4}$$

$N_F=N/L_F$ designates the total number of frames in X. It can be shown that each frame sequence or vector X_i can be spanned to a vector space formed by the orthonormal vectors $\{\phi_{ik}\}$.

Let the real orthonormal vectors be the columns of a transposed transformation matrix (Φ_i^T),

$$\Phi_i^T=\left[\phi_{i1}\quad \phi_{i2}\quad \ldots\quad \phi_{iL_F}\right] \tag{5}$$

it is evident that

$$X_i=\Phi_i^T C_i \tag{6}$$

where

$$C_i^T=\left[c_1\quad c_2\quad \ldots\quad c_{L_F}\right] \tag{7}$$

From the property

$$\Phi_i^T=\Phi_i^{-1} \tag{8}$$

that identifies Φ_i as an orthonormal matrix, the equations given below are obtained respectively.

$$\Phi_i X_i=\Phi_i\Phi_i^{-1}C_i \tag{9}$$

$$C_i=\Phi_i X_i \tag{10}$$

Thus, X_i can be written as a weighted sum of these orthonormal vectors

$$X_i=\sum_{k=1}^{L_F} c_k\phi_{ik}\ ,\quad k=1,2,3,\ldots,L_F \tag{11}$$

and the coefficients of the sub-frames are obtained as

$$c_k=\phi_{ik}^T X_i\ ,\quad k=1,2,3,\ldots,L_F \tag{12}$$

Let $X_{il}=\sum_{k=1}^{l} c_k\phi_{ik}$ be the truncated version of (11) such that $1\le l\le L_F$ [10]. It is noted that if $l=L_F$ then $X_i=X_{il}$. In this case, the approximation expression (X_{il}), and the approximation error (ε_l) are given by

$$X_{il}=\sum_{k=1}^{l} c_k\phi_{ik} \tag{13}$$

$$\varepsilon_l=X_i-X_{il}=\sum_{k=l}^{L_F} c_k\phi_{ik} \tag{14}$$

In this equation, ϕ_{ik} are determined by minimizing the expected value of the error vector with respect to ϕ_{ik} in the LMS sense**.** The above-mentioned LMS process results in the following eigenvalue problem. Eventually ϕ_{ik} are computed as the eigenvectors of the correlation matrix R_i of the sub-frame sequence X_i. By using orthonormality condition, the LMS error is given by

$$\varepsilon_l\varepsilon_l^T=\sum_{k=l}^{L_F} {c_k}^2 \tag{15}$$

Let J_l designate the expected value of the total squared error $\varepsilon_l\varepsilon_l^T$. Then,

$$J_l=E\left\{\varepsilon_l{\varepsilon_l}^T\right\}=\sum_{k=l}^{L_F} E\left\{{c_k}^2\right\} \tag{16}$$

$$E\left\{c_k^2\right\}=E\left\{\phi_{ik}^T(X_i^T X_i)\phi_{ik}\right\}=\phi_{ik}^T R_i\phi_{ik} \tag{17}$$

In order to obtain optimum transform, it is desired to find ϕ_{ik} that minimizes J_l for a given l, subject to the orthonormality constraint. Using Lagrangian multipliers,λ_k, we minimize

$$J=\sum_{k=l}^{L_F}\left[\phi_{ik}^T R_i\phi_{ik}-\lambda_k(\phi_{ik}^T\phi_{ik}-1)\right] \tag{18}$$

Taking the gradient of (18) with respect to ϕ_{ik}

$$\frac{\partial J}{\partial\phi_{ik}}=\frac{\partial}{\partial\phi_{ik}}\left[\sum_{k=l}^{L_F}\left[\phi_{ik}^T R_i\phi_{ik}-\lambda_k(\phi_{ik}^T\phi_{ik}-1)\right]\right]=0 \tag{19}$$

$$2R_i\phi_{ik}-2\lambda_k\phi_{ik}=0 \tag{20}$$

$$R_i\phi_{ik}=\lambda_k\phi_{ik} \tag{21}$$

is obtained. The value of the minimized J_l or the LMS error is then

$$\sum_{k=l}^{L_F}\phi_{ik}^T(\lambda_k\phi_{ik})=\sum_{k=l}^{L_F}\lambda_k \tag{22}$$

It is straightforward to obtain the matrix R_i as

$$R_i=\begin{bmatrix} r_i(1) & r_i(2) & r_i(3) & \ldots & r_i(L_F)\\ r_i(2) & r_i(1) & r_i(2) & \ldots & r_i(L_F-1)\\ r_i(3) & r_i(2) & r_i(1) & \ldots & r_i(L_F-2)\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ r_i(L_F) & r_i(L_F-1) & r_i(L_F-2) & \ldots & r_i(1)\end{bmatrix} \tag{23}$$

where

$$r_i(d+1)=\frac{1}{L_F-d}\sum_{j=[(i-1)\cdot L_F+1]}^{[(iL_F)-1-d]} x_{j+1}x_{j+1+d}\ ,\ d=0,1,2,\ldots,L_F-1 \tag{24}$$

It should be noted that the R_i is real, symmetric, positive-semi definite, and toeplitz. Obviously, λ_{ik} and ϕ_{ik} are the eigenvalues and the eigenvectors of the problem under consideration. It is well known that the eigenvalues of R_i are also real, distinct, and non-negative. Moreover, the eigenvectors ϕ_{ik} are all orthonormal. Let

eigenvalues be sorted in descending order such that $(\lambda_{1i} \geq \lambda_{1i} \geq \lambda_{1i} \geq ... \geq \lambda_{L_F i})$ with corresponding eigenvectors. The total energy of the sub-frame i is then given by $X_i{}^T X_i$.

$$X_i{}^T X_i = \sum_{k=1}^{L_F} x_{ik}{}^2 = \sum_{k=1}^{L_F} c_{ik}{}^2 = \sum_{k=1}^{L_F} \lambda_{ik} \quad (25)$$

Equation (11) may be truncated by taking the first p principal components [11], which have the highest energy of the original signal such that

$$X_i \cong \sum_{k=1}^{p} c_k \phi_{ik} \quad (26)$$

The simplest form of (26) can be obtained by setting p=1. The eigenvector ϕ_{i1} is called Major Signature Vector (MSV). That is to say, the MSV, which has the highest energy in the LMS sense, may approximate each frame that belongs to the original speech signal. Thus, we write

$$X_i \cong c_1 \phi_{i1} \quad (27)$$

In this case, L_F must be selected in such a way almost all the maximum energy is captured within the first term of (11) and the rest are negligible. Hence, (27) follows. Once (27) is obtained, it can be converted to equality by means of an "envelope diagonal matrix E_i" for each sub-frame. Thus, X_i

$$X_i = C_i E_i \phi_{i1}, \quad e_{ir} = \frac{x_{ir}}{C_i \phi_{i1r}}, \quad (r = 1,2,...,L_F) \quad (28)$$

In essence, the envelope terms e_{ir} of (28) somewhat absorbs the energy of the terms eliminated due to truncation of (10).

In this research work, many speech signals were examined and thousands MSV and envelope sequences were generated. It was observed that patterns obtained by plotting $e_i(n)$ (e_{ir} versus sub-frame index-n=1,2,3,…,L_F) and $\phi_{i1}(n)$ (ϕ_{i1r} versus sub-frame index-n=1,2,3,…,L_F) exhibit repetitive similarities.

It is deduced that these similar patterns are obtained due to the quasi-stationary behavior of the speech signals. In this case, one can eliminate the similar patterns obtained from many different experiments and thus form the so-called PSEVS or Banks with unique or one kind of patterns.

2.3 Elimination of Similar Patterns

In order to eliminate similar signature and envelope sequences, Pearson Correlation Coefficients (PCCs) are calculated. Two sequences are almost identical if PCC>0.9. The signature vectors which have unique patterns are combined under the set named "Predefined Signature Vector Set-PSVS $\{ S_{n_s}(n)$; n_s=1,2,3,...,N_S}". The integer N_S designates the total number of sequences in PSVS. Similarly, reduced envelope sequences are combined under the set called "Predefined Envelope Vector Set-PEVS $\{ E_{n_e}(n)$; n_e=1,2,3,...,N_E}". The integer N_E designates the total number of unique sequences in PEVS.

2.4 Reconstruction Algorithm

Step1: Divide X into the sub-frames X_i.

Step2a: For each sub-frame i pull an appropriate signature vector S_R such that the distance or the total error $\delta = \|X_i - C_R S_R\|^2$ is minimum for all $R = 1,2,3,..., N_S$.

Step2b: Store the index number R that refers to S_R ($X_i \approx C_R S_R$).

Step3a: Pull the appropriate E_K such that the error is further minimized for all $K = 1,2,3,..., N_E$. $\delta_K = \min\left\{\|X_i - C_R E_K S_R\|^2\right\}$.

Step3b: Store the index number K that refers to E_K. It should be noted that at the end of this step, the best S_R and the best E_K are found by appropriate selections. Hence, the sub-frame X_i is best described in terms of the patterns of E_K and S_R. i.e. $X_i \approx E_K S_R$.

Step4: Having fixed E_K and S_R, compute the new frame coefficient C_i in order to find the global minimum of the error $\delta_{Global} = \min\{\|X_i - C_i E_K S_R\|^2\}$ and store it. At this step, the sub-frame vector is approximated as $X_i \cong C_i E_K S_R$..

Step5: Repeat the above steps for each frame to reconstruct the original speech signal X(n).

3. IMPLEMENTATION OF THE NEW METHOD

In order to obtain the initial results, first the PSEVS were generated employing the proposed Algorithm. In this process, speech databases given by IPA Handbook [12] were utilized. PSEVS were constructed based on the combination of the words and sentences that comprise phonetics properties (vowels, consonants, tones, stress, conventions etc.) of different languages [12]. By investigating different frame length, optimum compression rate, which in term provides reasonable hearing quality, was found. The hearing quality was determined based on Mean Opinion Score (MOS) tests. In the proposed method, once PSVS and PEVS were generated, then any speech signal in any language were modeled frame-by-frame pulling the appropriate signature and envelope vectors from PSVS and PEVS using the new algorithm. The new algorithm was tested for 120 speech signals uttered by 5 female and 5 male speakers. 40 listeners are recruited in the speech quality assessments. Test words or sentences were taken from the IPA Handbook narratives that are not used in the construction stage of the PSEVS, OGI and TIMIT database [13]. Compression Rates vs L_F and Speech coding at different bit rates are shown in Table 1 and Fig.1. Comparison of the ADPCM and the new method by means of ACR-MOS of the speech samples uttered by female speakers in five languages and different bit rates is given in Fig.2. Furthermore objective test results summarized in Table 2. indicate that the new method offers almost the same performance of toll quality speech coding at 8Kbps bit rate while ADPCM exhibits the same performance at 16Kbps. It is understood that from the objective test results, the coder which is implemented by the new method at 4.5Kbps provides communication speech quality and the speech quality in the case of 2.31Kbps is slightly inferior to that of 4.5Kbps bit rate.

Table 1. Bit allocation table in different size of the PSEVS

L_F	Coded Parameter	Number of Vector	Bits /Frame	BR [Kbps]	CR
16	PSV	2048	11	16	4 : 1
	PEV	57422	16		
	FSC	-	5		
32	PSV	4096	12	8	8 : 1
	PEV	32768	15		
	FSC	-	5		
64	PSV	16384	14	4.5	14.22 : 1
	PEV	100992	17		
	FSC	-	5		
128	PSV	32768	15	2.31	27.52 : 1
	PEV	131072	17		
	FSC	-	5		

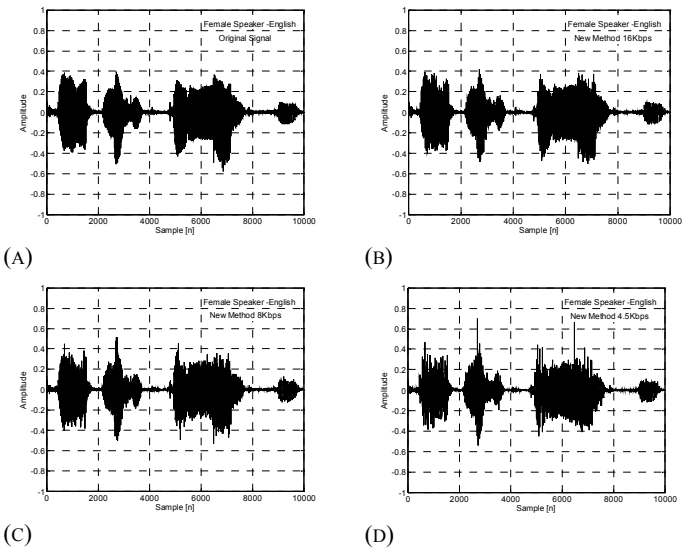

Fig. 1. (A) An original speech sample from the English female speaker database. Reconstruction (coding) of the speech sample in 16Kbps (B) 8Kbps (C) and 4.5Kbps (D) bit rates respectively.

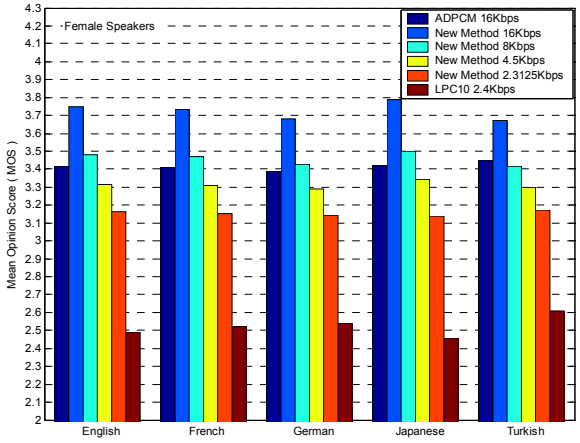

Fig. 2. Comparison of the ADPCM and the new method (MOS)

Table 2. Comparison of the objective test results (SNRseg) for female speakers

Language	Num. of Audio	SNRseg [dB]				
		ADPCM	New Method			
		16Kbps	16Kbps	8Kbps	4.5Kbps	2.31Kbps
English	12	7,4289	12,1969	8,0807	6,3536	4,9759
French	12	7,4396	12,0518	7,9134	6,0325	4,7153
German	12	6,6886	11,2053	7,3987	5,5582	4,3719
Japanese	12	11,1795	14,4533	10,2210	8,0923	6,9182
Turkish	12	7,6134	10,8320	7,1266	5,5262	4,4702
Compression Rate		4:1	4:1	8:1	14.22:1	27.52:1

4. CONCLUSIONS

In this paper, a new method is presented to synthesize or reconstruct the speech signals frame by frame by means of so-called PSEVS. The proposed signal representation method may suggest a new speech coding technique. In this coding scheme, the PSEVS are stored in each communication node and transmission of the speech is then achieved. Therefore, in the future research works, we wish to increase the computational efficiency to generate frame coefficients and to identify the indices of the predefined signature and envelope vectors on a proper DSP hardware also, we intend to comeup with a better schemes to connect the sub-frames in the course of synthesis to reduce the background noise. It is expected that the proposed algorithm may further be developed to handle some widely used speech processing applications such as speech recognition, speaker identification, language modeling, speech to text and word spotting. We also continued our effort in the area of speech recognition, and more specifically in word-specific predefined signature and envelope vectors.

REFERENCES

[1] S. Spanias, “Speech Coding: A tutorial review”, Proceedings of the IEEE, vol. 82, No. 10, 1994.

[2] Karaş, A. M., Yarman, B. S., “A new approach for representing discrete signal waveforms via private signature base sequences”, Proceedings of ECCTD'95 12th European Conference on Circuit Theory and Design, Istanbul, Turkey, August 27-31, 1995, pp. 875-878.

[3] A. M. Karaş, B. S. Yarman, “A new method for the compression of ECG signals: The Yar-Kar Method”, Proceedings of ICSPAT’97, San Francisco, USA, September 14-17, 1997.

[4] Akdeniz, R., Yarman, B. S., “Turkish speech coding by signature base sequences”, Proceedings of ICSPAT’98, Toronto, Canada, September 13-16, 1998, pp. 1291-1294.

[5] Ü. Güz, B. S. Yarman, H. Gürkan, “A new method to represent speech signals via predefined functional bases”, Proceedings of ECCTD’01, European Conference on Circuit Theory and Design, Espoo, Finland, August 28-31, 2001, vol. II, pp. 5-8.

[6] Ü. Güz, “A new approach in the determination of optimum signature base functions for Turkish speech”, Ph.D. thesis (Advisor: Prof. B. S. Yarman), Istanbul University, Inst. of Sci., Dept. of Electronics Eng., Istanbul, Turkey, Feb., 2002.

[7] B. S. Yarman, H. Gürkan, Ü. Güz, B. Aygün, “A new modeling method of the ECG signals based on the use of an optimized predefined functional database”, Acta Cardiologica-Int. Journal of Cardiology, vol. 58(3), pp. 273-275, 2003.

[8] Güz, Ü., Gürkan, H., Yarman, B. S., “A novel method to represent the speech signals by using language and speaker independent predefined functions sets”, Proceedings of ISCAS2004 International Symposium on Circuits and Systems, Vancouver, Canada, May 23-26, 2004, vol. III, pp. 457-460.

[9] Güz, Ü., Gürkan, H., Yarman, B. S., “A new speech signal modeling and word recognition method by using signature and envelope feature spaces”, Proceedings of ECCTD’03 European Conference on Circuit Theory and Design, Cracow, Poland, September 1-4, 2003, vol. III, pp. 161-164.

[10] A. N. Akansu, R. A. Haddad, “Multiresolution signal decomposition, transforms, subbands, wavelets”, Academic Press, Inc., San Diego, 1992.

[11] I. T. Jolliffe, “Principal component analysis”, Springer Series in Statistics. Springer-Verlag, New York Inc., 1933.

[12] IPA Handbook (Handbook of the international phonetic association a guide to the use of the international phonetic alphabet), Cambridge University Press, July 1999.

[13] OGI Multi-Language Telephone Speech Corpus, CD-ROM, Linguistic Data Consortium.