Published in:
Cybernetics and Information Technologies, vol.1, No.2, 2001, pp.19-29.

Clipped LPC Cepstrum and Its Application to Text-Independent Speaker Identification

Atanas Ouzounov
Institute of Information Technologies
Bulgarian Academy of Sciences
Acad. G. Bonchev Street bl.29A
Sofia 1113, Bulgaria
E-mail:atanas@iinf.bas.bg

Abstract: *A new modification of the LPC cepstrum of speech signals called Clipped LPC (CLPC) cepstrum is proposed. In the CLPC cepstrum is reduced the influence of the low-level LPC spectrum's regions. Three LPC cepstrums as features in a text-independent speaker identification task were evaluated using reading text in Bulgarian language collected over noisy telephone lines. These cepstrums are standard LPC cepstrum, CLPC cepstrum and OSALPC cepstrum. As experimental results shown, the proposed cepstrum achieves better results than both LPC and OSALPC cepstrums in this task.*

Keywords: *Speaker identification, Group delay spectrum, LPC cepstrum, Algebraic approach.*

1. Introduction

The performance of the existing speaker recognition systems degrades rapidly when training and testing cannot be done in the same ambient conditions. Ones of the most common representations in these systems are Linear Predictive Coding (LPC) cepstral-based parameters. It is well known that LPC cepstral parameters might lead to a poor recognition rate in the noisy environment. One of the solutions to cope the noise problem is to find cepstral representations, which are resistant to the noise corruption [8].

In [10] it is shown that the influence of the noise on the logarithmic power spectrum of the speech signal is essential in the low-level domains. This influence leads, during the recognition process, to instability of the values of the used spectral distance and this is detrimental for speech recognition. Therefore, it is reasonable to propose that if the influence of these domains were suppressed in the LPC cepstrum, it would bring to lower variance of cepstral coefficients of noisy speech. This could be made by using the properties of Group Delay Spectrum (GDS) [1, 2, 3, 4].

It is known that the positive regions of the smoothed GDS approximate the formants and the negative regions - the spectral valleys. To reduce the influence of spectral valleys in LPC cepstrum it is need to remove (or to decrease) the GDS in selected negative regions [2, 3, 4].

In this paper, we are proposing a new modification of the LPC cepstrum analytically obtained by use of a polynomial approximation of the GDS's negative values reduction function. We named the obtained cepstrum a Clipped LPC (CLPC) cepstrum and suggested simplified formula for its calculation. Some preliminary experimental results for the clipped real cepstrum are presented in an earlier work of the author [9].

We evaluate three LPC cepstrums as features in a text-independent speaker identification task with phrases in Bulgarian language collected over noisy telephone channels. These cepstrums are standard LPC cepstrum, CLPC cepstrum and the One-Sided Autocorrelation Linear Predictive Coding (OSALPC) cepstrum [7]. As a classification scheme, we use an algebraic approach with arithmetic-geometric sphericity measure [5, 6].

2. Group delay spectrum

If $X(\Omega)$ is the Fourier transform (FT) of the minimum phase signal $x(i)$, $c(n)$ are the cepstral coefficients of $x(i)$ and $X(\Omega) = |X(\Omega)| \exp(j\theta_x(\Omega))$, then $\ln X(\Omega)$ can be represented as [3]

$$\ln X(\Omega) = 0.5c(0) + \sum_{n=1}^{\infty} c(n) \exp(-j\Omega n) . \quad (1)$$

The unwrapped phase function is

$$\theta(\Omega) = \theta_v(\Omega) + 2\pi\lambda(\Omega) = \sum_{n=1}^{\infty} c(n) \sin(\Omega n) \quad (2)$$

where $\theta_v(\Omega)$ is wrapped phase function and $\lambda(\Omega)$ is an integer such that $\theta(\Omega)$ is a continuous function of Ω [3].

The group delay spectrum $\tau(\Omega)$ is defined as negative derivative of the unwrapped phase function with respect to Ω [3]. For minimum phase signal, $\tau(\Omega)$ is

$$\tau(\Omega) = -\frac{d\theta(\Omega)}{d\Omega} = \sum_{n=1}^{\infty} nc(n) \cos(\Omega n) \quad (3)$$

For mixed phase signal we will define two GDSs - $\tau_m(\Omega)$ and $\tau_p(\Omega)$:

$$\tau_m(\Omega) = \sum_{n=1}^{\infty} nc_m(n) \cos(\Omega n) ; \qquad \tau_p(\Omega) = \sum_{n=1}^{\infty} nc_p(n) \cos(\Omega n) , \quad (4)$$

where $c_m(n)$ and $c_p(n)$ are the cepstral coefficients of the minimum phase equivalent signals derived from spectral magnitude and phase of $X(\Omega)$ respectively [3].

3. Noisy speech processing using group delay spectrum properties

The additive and high-resolution properties of the GDS allow its domains to be separately processed. In this work, the negative domains of the GDS are of interest, because they correspond to low level logarithmic power spectrum [2, 10].

It was shown in [2] that the substitution of $\tau_m(\Omega)$ in the negative domains by their mean values leads to a reduction of the noise influence in the LPC spectrum. This approach requires calculation of two GDSs $\tau_m(\Omega)$ - $\tau_m^u(\Omega)$ and $\tau_m^s(\Omega)$. $\tau_m^u(\Omega)$ is an unsmoothed GDS and for its calculation are used all cepstral coefficients (typically their number is equal to the size of cos transform). $\tau_m^s(\Omega)$ is a smoothed GDS and for its calculation are used only part of coefficients (typically the first 10 or 20). The negative values' regions are determined in smoothed GDS. In the unsmoothed GDS, the levels in these frequency regions are substituted with their mean values. The inverse cos transform is then used to obtain the modified cepstrum $c_m^r(\Omega)$ in which the influence of low-level spectral regions is reduced. The reconstruction of the autocorrelation function is done by processing of the cepstrum $c_m^r(\Omega)$ via an inverse homomorphic system. The values of the autocorrelation function are used by Levinson-Durbin algorithm to produce the linear prediction coefficients. From them the LPC cepstrum is calculated.

This approach is effective but it is very complicated and it does not allow analytical solution. Our aim is to reduce (not as effective as in [2]) the variance of the LPC cepstral coefficients for noisy speech using a similar idea and in a more simple way.

4. Clipped LPC cepstrum

In the paper we will use the GDS $\tau(\Omega)$ in (3) obtained by LPC cepstrum. In the case we will process a minimum phase part of speech signal consists of information only for poles of speech spectrum ($c_m(n)$ in (4) consists of information for zeros and poles). It is well known that in all-pole model of linear prediction analysis only the spectral peaks are well represented. In this model, the low-level spectral regions are approximated with insufficient accuracy and their variability (caused by a bad approximation or a noise influence) is one of the sources of the LPC cepstrum variability. We suppose that if we smooth and reduce the spectrum values in these regions then we can decrease the LPC cepstral coefficients' variance.

Let $c(n)$ is the LPC cepstrum of the analyzed speech signal, the GDS $\tau(\Omega)$ is defined as in (3) and the smoothed GDS obtained by using the first M LPC cepstral coefficients in (3) is noted as $\tau_s(\Omega)$. The GDS formed by positive terms of $\tau_s(\Omega)$ is noted as $\tau_s^+(\Omega)$ and

$$\tau_s^+(\Omega) = W[\tau_s(\Omega)] = \tau_s(\Omega)Y[\tau_s(\Omega)] + GY[-\tau_s(\Omega)], \quad (5)$$

where $W[.]$ is a negative values' clipping function, G is a constant and $Y[.]$ is the Heaviside step function. The LPC cepstrum corresponding to $\tau_s^+(\Omega)$ is noted as $c^+(k)$ and according to (3) and (5), for $k > 0$, it is

$$c^+(k) = \frac{2}{k\pi}\int_0^\pi \tau_s^+(\Omega)\cos(\Omega k)d\Omega = \frac{2}{k\pi}\int_0^\pi \{\tau_s(\Omega)Y[\tau_s(\Omega)] + GY[-\tau_s(\Omega)]\}\cos(\Omega k)d\Omega . \quad (6)$$

The direct solution of (6) by using of an analytical form of $W[.]$ is a very difficult task. Therefore, we are proposing to use in (6) a polynomial approximation of the clipping function $W[.]$. This kind of approximation (performed by Least Squares Method (LSM)) allowed us to evaluate analytically the integral. We note the approximation of $W[.]$ as $W_P[.]$, and the GDS obtained by $W_P[.]$ as $\tau_s^{+p}(\Omega)$. We call the LPC cepstrum corresponding to $\tau_s^{+p}(\Omega)$ Clipped LPC (CLPC) cepstrum and note it as $c_c(k)$. According to (3) and (6), we have

$$c_c(k) = \frac{2}{k\pi}\int_0^\pi \{\sum_{r=0}^{R} b_r[\sum_{n=0}^{M} nc(n)\cos(\Omega n)]^r\}\cos(\Omega k)d\Omega , k > 0, \quad (7)$$

where R is the order and b_r - coefficients of the polynomial.

The approximation of $W[.]$ is performed for $\tau_s(\Omega)$ between $\tau_s(\Omega)_{\min}$ and $\tau_s(\Omega)_{\max}$. The values of $\tau_s(\Omega)_{\min}$ and $\tau_s(\Omega)_{\max}$ depend on the amplitude range of $\tau_s(\Omega)$. This range is determined by few factors: speech signal properties, an all-pole model order and a number of cepstral coefficients. The experiments revealed, that when we used autocorrelation method with model order $P \le F_s + 10$, where F_s is sampling frequency in kHz and number of LPC cepstral coefficients $M \le 3P$, then the inequality $\tau_s(\Omega)_{\min} \le \tau_s(\Omega) \le \tau_s(\Omega)_{\max}$, where $\tau_s(\Omega)_{\min}$ = -30 and $\tau_s(\Omega)_{\max}$ = 30, is always fulfilled.

It was found that $c_c(k)$ in (7) could be efficiently evaluated when $R \le 3$. For R =3 the used polynomial coefficients are $b_0 = 0$, b_1 = 0.72734, b_2 = 0.02052, b_3 = -0.00013. They are obtained after correction *ad hoc* on LSM derived coefficients. Through this correction the positive values in $\tau_s^{+p}(\Omega)$ are additionally enhanced.

We obtain CLPC cepstrum $c_c(k)$ in (7) as a sum of three terms: $c_{c1}(k)$, $c_{c2}(k)$ and $c_{c3}(k)$. The (real) LPC cepstrum is $c(n)=0$ for $n<0$ [1]. In this case the terms of the CLPC cepstrum $c_c(k), k=1,\cdots,M_c$, $M_c \le M$, where M_c is the number of the CLPC cepstral coefficients, are

$$c_{c1}(k) = b_1 c(k); \tag{8}$$

$$c_{c2}(k) = (b_2 / k)[R(k) + 0.5R^{-}(k)]; \tag{9}$$

$$c_{c3}(k) = 0.25(b_3 / k)\sum_{n=1}^{M-k} nc(n)\sum_{v}[R(v) + R^{-}(v)], \tag{10}$$

where $v = -n-k, -n+k, n-k, n+k$ and $R(.)$, $R^{-}(.)$ are correlation functions of the index-weighted LPC cepstrum of the form

$$R(q) = \sum_{l=1}^{M-q} lc(l)(l+q)c(l+q); \qquad R^{-}(q) = \sum_{l=1}^{q} lc(l)(-l+q)c(-l+q) \tag{11}$$

We ignore $c_{c3}(k)$ in (10) ($b_3 << b_2$) and according to (8) and (9), we have the following simplified formula for $c_c(k), 1 \le k \le M_c$

$$c_c(k) = b_1 c(k) + (b_2 / k)\sum_{n=1}^{M-k} nc(n)(n+k)c(n+k) + (0.5b_2 / k)\sum_{n=1}^{k-1} nc(n)(k-n)c(k-n), \tag{12}$$

5. OSALPC cepstrum

It is known that the autocorrelation sequence preserves the all-pole properties of original time sequence. The OSALPC cepstrum is obtained by Autoregressive (AR) modeling of the causal part of the autocorrelation sequence. In some speaker identification research works has shown that in additive white noise experiments and in noisy car environment this cepstrum produces lower recognition error than the standard LPC cepstrum [7]. The description of the algorithm for OSALPC cepstrum calculation is given bellow [7]:

- -from speech segment of length *N* samples the AutoCorrelation Function (ACF) with lags *M=N/2* is computed;
- -the 0^{th} lag is set to zero *R(0) = 0*;
- -the Hamming window from *m=0* to *M* is applied to one-sided ACF obtained in the previous two steps;
- -the first *p+1* autocorrelation lags of this sequence are computed using classical biased estimator;
- -these values are used by Levinson-Durbin algorithm to produce the AR parameters;
- -the cepstral coefficients are recurrently obtained from those AR parameters.

6. Experiments

6.1. Normalized variance estimation

In this experiment we are evaluated the normalized variances of the LPC, CLPC and OSALPC cepstral coefficients for telephone speech corrupted by an additive white noise. The speech signal is sampled at 8 kHz. A 14^{th} order autocorrelation analysis without preemphasis is used. The variance of the each cepstral coefficient is measured over 3997 frames (frames of 32 ms at 8 ms rate). The speech is corrupted by additive white Gaussian noise with segmental signal-to-noise ratio (SNR) 20, 10 and 0 dB. The normalised variance is $V_N(c(n)) = V(c(n))/V(c(1))$, where $V_N(c(n))$ is the variance of n^{th} cepstral coefficient and $V_N(c(1))$ is the variance of the first coefficient. In Figures 1, 2 and 3 are shown the values for $V_N(.)$ respectively for standard LPC, CLPC and OSALPC cepstrums and different SNRs.

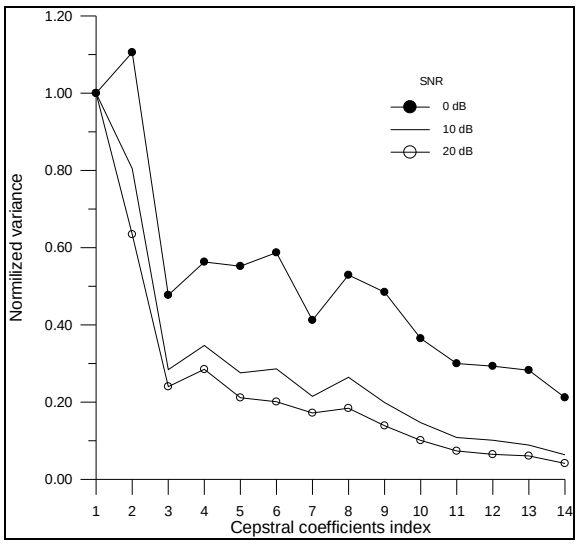

Fig.1. LPC cepstrum, SNR 20, 10 and 0 dB

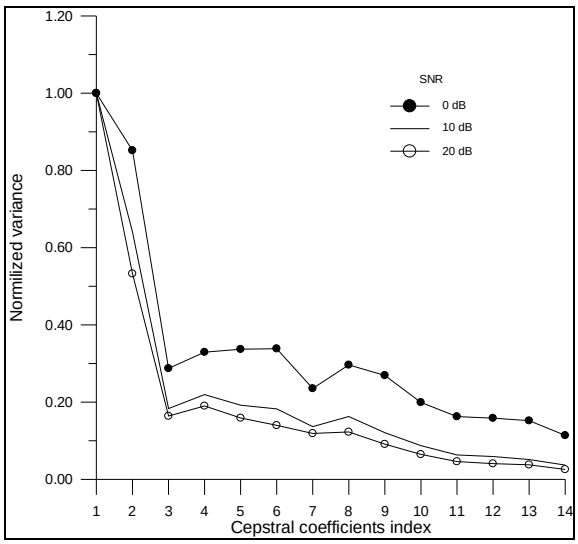

Fig.2. Clipped LPC cepstrum, SNR 20, 10 and 0 dB

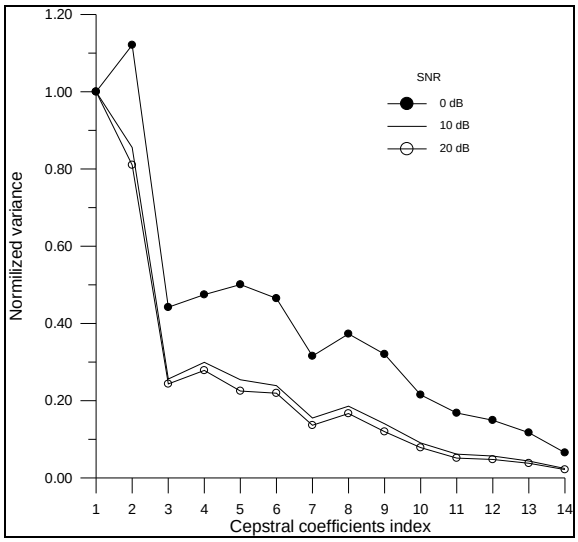

Fig.3. OSALPC cepstrum, SNR 20, 10 and 0 dB

On the base results gained from the experiment, we can make the following conclusions:

-at the same SNR the CLPC and the OSALPC cepstrums produce lower than the standard LPC cepstrum normalized variance for all coefficients;

-the normalized variance for the CLPC cepstrum is less than this one for OSALPC cepstrum especially for SNR 0 dB and cepstral index from 2 to 12;

The CLPC cepstrum is obtained by non-linear transformation of the GDS that acts on spectrum level. We suppose that CLPC cepstrum will be more effective for voiced frames because for them the non-linear spectrum level reducing is more significant. Probably for unvoiced frames, the similar approach would be lead to negligible effect.

6.2. Speaker identification experiments

6.2.1. Speech database

The analysed speech (reading text in Bulgarian language) is recorded over dialed-up analogues telephone lines in the city of Sofia. The reading text lasts about 40 seconds and was read at least 2 times from each of 12 speakers (male). Every repetition is recorded from different

telephone call. The speech database for each speaker is divided in two parts - for training and for recognition. These two parts are of different linguistic contents and are obtained from different telephone calls. We did not accomplish preliminary selection of the speech data for training and testing based on its noisy characteristics.

6.2.2. Speaker identification technique

Many techniques have demonstrated good text-independent speaker identification performance. For the current task we have selected one that is easy to implement and has in the same time high computational efficiently. That is an algebraic approach with arithmetic-geometric sphericity measure [5, 6].

Algebraic approaches are well-known techniques in text-independent speaker recognition tasks. They are based on an estimation of the covariance matrix of the sequence of the parametric vectors of speech data. The description of the selected technique is a given bellow [5].

From the sampled speech signal, we compute parametric presentation - sequence of N vectors in p-dimensional space. Let $\{X_n\}$ is the vector sequence and $n=1,...,N$. The covariance matrix X of $\{X_n\}$ is obtained as [5]:

$$X = \frac{1}{N}\sum_{n=1}^{N} X_n X_n^T - m_x m_x^T; \tag{13}$$

$$m_x = \frac{1}{N}\sum_{n=1}^{N} X_n, \tag{14}$$

where - m_x is the mean vector of $\{X_n\}$ and T is matrix transposition. The covariance matrix of a test sequence $\{Y_l\}, l = 1,..., L$ is denoted as Y.

There are measures that can be considered as different estimations of the similarity of two covariance matrices [5]. They belong to the distances that could be defined using only the eigenvalues of the product YX^{-1}. Let the p-ordered eigenvalues of the matrix YX^{-1} are denoted as $\{\lambda_i\}$ and three particular functions of $\{\lambda_i\}$ are defined [5]

$$A(\lambda_1,...,\lambda_p) = \frac{1}{p}\sum_{i=1}^{p}\lambda_i; \tag{15}$$

$$G(\lambda_1,...,\lambda_p) = \sqrt[p]{\prod_{i=1}^{p}\lambda_i}; \tag{16}$$

$$H(\lambda_1,...,\lambda_p) = p\left(\sum_{i=1}^{p}\frac{1}{\lambda_i}\right)^{-1} \tag{17}$$

Functions A, G and H are respectively arithmetic, geometric and harmonic means of the eigenvalues $\{\lambda_i\}$. The swapping of the matrices X and Y leads to transformation of A into H^{-1}, G into G^{-1} and H into A^{-1}.

The arithmetic-geometric sphericity measure $\mu_{AG}(X,Y)$ [5] is

$$\mu_{AG}(X,Y) = \log\left(\frac{A}{G}\right) \tag{18}$$

The measure equals zero if $A=G$, that is when all eigenvalues are equal - i.e. X and Y are proportional. The arithmetic-geometric sphericity measure is non-symmetric. If we take into account the length of data for X and Y estimation then the general symmetric measure $\mu^{sym}(X,Y)$ is [5]

$$\mu^{sym}(X,Y) = \alpha_{LN}.\mu_{AG}(X,Y) + \alpha_{NL}.\mu_{AG}(Y,X) \tag{19}$$

where $\alpha_{LN} + \alpha_{NL} = 1$.

The weighting functions α_{LN} and α_{NL} are:

$$\alpha_{LN} = \frac{L}{M+L};\ \alpha_{NL} = \frac{N}{L+N} \tag{20}$$

In the training mode, one reference is used per speaker, which is the covariance matrix of cepstral vectors of training data. In testing mode we calculate the arithmetic-geometric measures $\mu_s^{sym}(X,Y)$ in (19) between covariance matrix *Y* obtained from input speech sequence from speaker with unknown identity and reference matrices X_s of all speakers, $s = 1,\cdots,S$. Then we make decision based on minimum distance rule. No additional threshold is used.

6.2.3. Experiments and results

The speech signal is digitized at 8 kHz on 16 bits, after low-pass filtering at 4 kHz. Preemphasis is not applied. Hamming windowing frames with 32 ms length are used, with frame rate of 8 ms. A 14 th order autocorrelation analysis is carried out. Each frame has been converted into 14th order cepstral vector - standard LPC, CLPC and OSALPC. In our study the pauses has been removed by energy threshold and their lengths are not included further in processing data length.

We examined the recognition rate as a function of training data. The data pool has been clustered into three different lengths of data for training - 4, 8 and 12 seconds. In testing mode, we processed supra frames with length of 4 seconds and frame rate of 0.5 seconds. The number of tests is 265 and the identification error is averaged over 12 speakers. For this case, the experimental results are shown in Table I. If we exclude only from testing data the speakers with the worst recognition rate then the number of tests is 216 and the results are shown in Table II.

Table I. Identification error in percentage averaged over 12 speakers

	Test - 4 seconds		
Training data in seconds	LPC	CLPC	OSALPC
4	54.33	51.69	56.22
8	45.66	40.00	48.30
12	36.98	32.83	45.66

Table II. Identification error in percentage averaged over 10 speakers (without the worst 2 speakers)

	Test - 4 seconds		
Training data in seconds	LPC	CLPC	OSALPC
4	43.98	40.74	46.29
8	33.33	26.38	36.57
12	22.68	17.59	33.33

7. Discussion and conclusions

To obtain the results shown in Table I, two additional experiments have been done. In the first one, we calculated the CLPC cepstrum for all segments in speech data (without pauses). The recognition rate in few of the tests was close to the LPC one, but for the rest was worse. In the second experiment, we set an additional energy threshold to produce voiced segments selection. Then we used CLPC cepstrum only for voiced segments and the LPC one for the rest. In that case, we obtained the results already shown in Table I. The detailed analysis of CLPC cepstrum properties shows that the recognition rate depends on the order *M* of the LPC cepstrum used in (12). We observed that the best results were produced when $M \approx 3M_c$ in (12). In our case $M_c = 14$ and $M = 50$.

We expected from the OSALPC cepstrum to demonstrate a better recognition rate. We analyzed the OSALPC cepstrum behaviour for each speaker and noticed that for some cases this method gives the best results and vice versa for others. The excellent results were obtained in the cases when the training speech was clean and the test one contained only background noise (similar to the white noise). In the rest cases - when the noise is pulse or there is a crosstalk and harmonic noise or the training data is noisy but the testing – clean speech, the OSALPC cepstrum produced unsatisfactory results.

The analysis of experimental results revealed that there are two speakers, which cannot be identified correctly in all tests and for all cepstrums. For these cases, it appears that the combination of the selected cepstrums and identification technique has not enough discrimination capability to achieve the correct recognition. We excluded these two speakers from the test data and we obtained a significant reduction of the identification error, as it is shown in Table II.

The results from our study demonstrate that this cepstrum modification is effective in real telephone speech identification tasks. This motivates us in our forthcoming work to examine the CLPC cepstrum as feature in others speaker identification techniques.

REFERENCES

1. Singer, H., T. Umezaki, and F. Itakura, Low Bit Quantization of the Smoothed Group Delay Spectrum for Speech Recognition, In Proceedings of the ICASSP, 1990, pp. 761-764.
2. Thomas, J. A., B. Yegnanarayana, R. Karinthi, and V. Venkateswar, Processing of Noisy Speech using Group Delay Functions, In Proceedings of the ICASSP, 1985, pp.720-723.
3. Yegnanarayana B., D. Saikia, and T. Krishnan, Significance of Group Delay Function in Signal Reconstruction from Spectral Magnitude or Phase, *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 32, No.3, 1984, pp.610-623.
4. Yegnanarayana B., K. Murthy, and H. Murthy, Applications of Group Delay Functions in Speech Processing, *The Journal of the Institution of Electronics and Telecommunication Engineers of India*, vol.34, No.1, 1988, pp.20-29.
5. Bimbot F., I. Margin, and L. Mathan, Second-Order Statistical Measures for Text-Independent Speaker Identification, *Speech Communication*, vol.17, No.1-2, 1995, pp.177-192.
6. Bimbot F. and L. Mathan, Second Order Statistical Measures for Text-Independent Speaker Identification, *Proceedings of ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, Switzerland, 1994, pp.51-54.
7. Hernando J., et al., Speaker Identification in Noisy Conditions using Linear Prediction of the One-Sided Autocorrelation Sequence, In Proceedings of the ICSLP'94, pp.1847-1850.

8. Mammone R., et al., Robust Speaker Recognition, *IEEE Signal Processing Magazine*, September 1996, pp.58-71.
9. Ouzounov A., Clipped Cepstrum - A Real Cepstrum of Speech Signals Robust to Additive Noise, *Problems of Engineering Cybernetics and Robotics*, vol.41, 1994, pp.19-26.
10. Takahashi K., Y. Matsumoto, H. Kobatake, Studies on Noisy Word Recognition, *Systems and Computers in Japan*, vol. 17, No. 5, 1986, pp. 1-7.