SPEECH WATERMARKING BASED ON ROBUST PRINCIPAL COMPONENT ANALYSIS AND FORMANT MANIPULATIONS

*Shengbei WANG, Weitao YUAN, Jianming WANG**

School of Computer Science
& Software Engineering
Tianjin Polytechnic University
Binshuixi Road, Xiqing District, Tianjin, China

Masashi UNOKI†

School Information Science
Japan Advanced Institute of Science
and Technology
1-1 Asahidai, Nomi, Ishikawa, Japan

ABSTRACT

This paper proposes a watermarking method for speech signals based on Robust Principal Component Analysis (RPCA) and formant manipulations. As the spectrogram of speech has a relatively sparse structure, the core information of speech is extracted into a sparse matrix using RPCA so that formants can be estimated with Linear Prediction (LP) more accurately even under noise/interferences, which significantly improves the robustness of proposed method. We investigate how the formants can be controlled and manipulated to make the watermarking method effective. Watermarks are embedded into speech by controlling the shape and power of formants using the stable and robust parameter, i.e., line spectral frequencies (LSFs). Evaluations regarding inaudibility and robustness are carried out and the results suggest that the proposed method can not only satisfy inaudibility but also provide good robustness against general processing and different speech codecs which is better than the other methods.

Index Terms— Robust principal component analysis, Linear prediction, Formant, Line spectral frequencies, Robustness

1. INTRODUCTION

Speech signal is an important information carrier in many social applications such as WeChat and GoogleTalk. However, modern digital technologies have put the security of speech at risk. Watermarking is a promising solution to protect speech signals. A general watermarking should be inaudible to human perception, blind for watermark extraction, and robust against signal processing/codecs. However, there is a trade-off among these competitive requirements, e.g., robustness is usually improved at the expense of inaudibility, and vice versa. Therefore, how to realize desired watermarking is still a challenging problem. This work focuses on exploring inaudible, blind, and robust speech watermarking.

There has been significant research into speech watermarking recent years. A typical category of watermarking focuses on exploring the characteristics of human auditory system (HAS) for inaudibility [1, 2]. For instance, watermarks can be embedded into the phase of speech based on fact that HAS is not sensitive to slight phase modifications [3, 4]. Quantization index modulation (QIM) [5, 6] based methods form another category of watermarking, where a lot of efforts have been devoted to selecting suitable features to balance inaudibility and robustness. Spread spectrum is a well-known technique which is widely employed for robust watermarking [7, 8, 9]. Aside from these categories, hybrid watermarking [10, 11, 12] has been verified to have superior performance in robustness since watermarks are doubly embedded which enables them to be reliably extracted. Despite these achievements, many existing methods cannot reach a balance between inaudibility and robustness. In particular, robustness against codecs is highly desired for speech watermarking while many methods are not completely robust against different speech codecs.

A common problem in watermarking field is that many methods can extract the watermarks in ideal situations (without noise/ interferences), but when there are noise/interferences in watermarked signal, the embedded watermarks will fail to extract which leads to weak robustness. We previously proposed two formant enhancement based watermarking methods [13, 14]. However, their robustness against speech codecs was not satisfactory, e.g., [13] was not robust against any speech codecs and [14] was not robust against G.729 at high capacities. This paper proposes a speech watermarking method based on robust principal component analysis (RPCA) and formant manipulations. RPCA is employed to extract the core information in speech so that formants can be estimated correctly even under interferences caused by speech processing and codecs. Watermarks are embedded into the formants of relatively low power by controlling line spectral frequencies (LSFs) to maintain the speech quality. The main contribution of this paper is that RPCA is introduced to watermarking for the first time and the introducing of RPCA can significantly attenuate the influence of various interferences in watermark extraction process which improves the robustness. The effectiveness of proposed method is demonstrated in the experiments.

2. PROPOSED METHOD

Linear Prediction (LP) is popular for separating the vocal tract and excitation information in the source-filter model of speech production. The LP coefficients derived from LP can provide important information of acoustic feature, i.e., formants. Nevertheless, when speech is smeared by interferences such as background noise and reverberation, the estimated LP envelope and formants will be much distorted. As the proposed method embeds watermarks into formants, it is necessary to make sure that formants could be correctly estimated even under interferences.

In general, speech varies significantly and continuously over time and its power concentrates on formants, thus the spectrogram of speech has a relatively sparse structure. Based on this fact, some

*Thanks to grant No. 2017KJ089, Natural Science Foundation of Tianjin (No. 17JCQNJC00100 and No. 16JCYBJC41500), and National Natural Science Foundation of China (No. 6137104 and No. 61602344) for funding.

†This work was also supported by a Grant-in-Aid for Scientic Research (B) (No. 17H01761) and I-D DATA foundation.

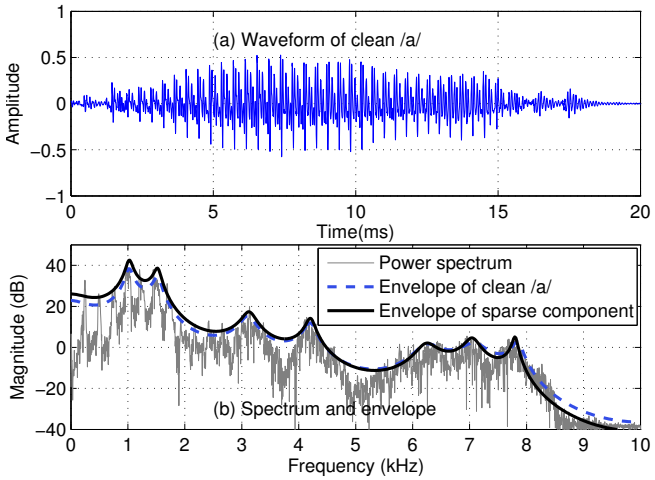

Fig. 1. Formant estimation for /a/ with and without RPCA.

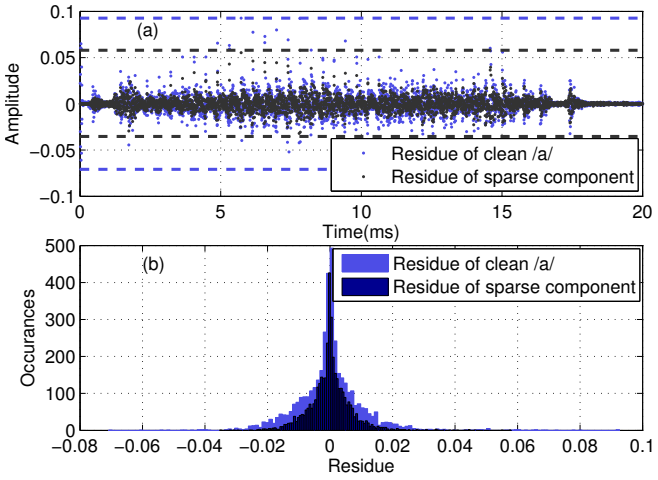

Fig. 2. Residue analysis for clean speech.

works successfully separated clean singing voice/speech from music accompaniment and interferences using RPCA [15] with the assumption that the singing voices/speech have a sparse structure and the music accompaniment and interferences have a low-rank structure. In watermarking field, common speech processing/codecs will unavoidably introduce interferences to speech. If we can separate out the core information of speech from interferences to reduce their effect, the robustness of watermarking will be improved. This paper employs RPCA to extract the core information of speech so that formant estimation could be stable and robust under these interferences.

2.1. Robust principal component analysis

RPCA as a convex problem is a matrix factorization algorithm which is quite attractive in various application fields [16]. Given an input matrix $X \in R_{n_1 \times n_2}$, RPCA decomposes it into a sparse matrix $S \in R_{n_1 \times n_2}$ and a low-rank matrix $L \in R_{n_1 \times n_2}$ by solving the following convex problem,

$$\begin{aligned} &\text{minimize}_{S,L} && \frac{\lambda_0}{\sqrt{\max(n_1, n_2)}} \|S\|_{l_1} + \|L\|_* \\ &\text{subject to} && X = S + L, \end{aligned} \tag{1}$$

where S (punished by the l_1-norm $\|\cdot\|_{l_1}$, i.e., the sum of absolute values of matrix entries) represents the sparse component of X, L (punished by the nuclear norm $\|\cdot\|_*$, i.e., the sum of singular values) represents the low-rank component of X, and $\lambda_0 > 0$ is a trade-off parameter to adjust the ratio between the sparse component S and low-rank component L. As suggested in [15], $\lambda_0 = 1$ is a good choice, but this paper adjusts λ_0 to a smaller value to relax the restrictions on S so that the main spectra of speech could be captured in sparse component S. The inexact Augmented Lagrange Multiplier (ALM) method [15, 16] is used to solve the RPCA problem and two matrices S and L can be obtained.

PRCA is operated on matrix, this paper transforms the clean/noisy speech into a matrix in the time-frequency (T-F) domain with short-time Fourier transform (STFT). The obtained matrix is a joint matrix composed of a sparse matrix of core information in speech and a low-rank matrix which contains the noise/interferences. The performance of RPCA is verified in next subsection.

2.2. Linear prediction analysis and formant estimation

LP analysis can provide the estimation of formants [17, 18]. A common representation of LP is expressed in Eq. (2), where p is the LP order, a_i are LP coefficients, $\hat{x}(n)$ is the predicted value, and $x(n-i)$ is the i-th previous values,

$$\hat{x}(n) = \sum_{i=1}^{p} a_i x(n-i). \tag{2}$$

The performance of formant estimation with and without RPCA in clean and noisy/interferences conditions is verified using voiced vowel /a/ (20 kHz, 16-bit). In RPCA, the spectrogram of speech is computed with window size of 256. The sparse component in speech is calculated using Eq. (1) where λ_0 is fixed at 0.2 to ensure the core information of speech is captured in sparse component. LP order for formant estimation is 16-th. Figure 1(b) shows the LP envelope and formant estimation results from clean /a/ and its sparse component. The formants estimated from sparse component is quite close to those from clean /a/, indicating that formant structure is well kept in sparse component. In Fig. 2(a), the LP residue of sparse component closely concentrates in a smaller numerical range than that of clean /a/ and its histogram in Fig. 2(b) is much thinner and higher, which means the residue of sparse component is much smaller. The mean value (MV), standard deviation (STD), and normalized mean value (NMV) of residues are calculated in Tab. 1 (line 2-3). We can confirm that for clean speech, formant structure can be well kept in sparse component.

In noisy condition, a Gaussian white noise at 20 dB is added to clean /a/. Formants are estimated from noisy speech and its sparse component respectively. The histogram of LP residues for noisy /a/, its sparse component, and clean /a/ are compared in Fig. 3. The residue of clean /a/ is the thinnest and highest and the LP residue of sparse component of noisy /a/ is similar with clean /a/, while the residue of noisy /a/ is much larger. The statistical data of LP residues for noisy /a/ and its sparse component are listed in Tab. 1 (line 4-5). It is found that when there are interferences in speech, formant estimation from the sparse component is more accurate.

2.3. Concept of watermarking

The estimated formants are originally expressed with LP coefficients. In this paper, LP coefficients are converted to LSFs

Table 1. Statistical analysis for residue.

Conditions	Conditions	MV	STD	NMV
Clean	/a/	0.007	0.011	0.013
	Sparse component	0.004	0.006	0.009
Noisy	/a/	0.637	0.788	0.041
	Sparse component	0.016	0.021	0.028

[19, 20, 21] for formant manipulations as they possess several properties: (i) LSFs are less sensitive to noise; (ii) the influences caused by LSF deviations can be limited to local spectra, which suggests that if LSFs are used to manipulate the formant for watermark embedding, the sound distortion introduced by watermarks can be minimized; (iii) LSFs are universal features in different speech codecs, watermarking implemented on LSFs is possible to survive from speech codecs to provide strong robustness. The relationship between LSFs and formants is that one formant can be controlled by two adjacent LSFs, and the closer two LSFs are, the sharper the formant is. Watermarks can be embedded into speech when LSFs are shifted for formant manipulations.

Embedding concept: Formants in low frequencies are important for speech quaility and speech recognition, thus this paper chooses formants in high frequencies which have relatively low power for watermark embedding. As shown in Fig. 4(a), suppose the last two formants of one speech frame are controlled by four LSFs, ϕ_a, ϕ_b, ϕ_c, and ϕ_d. The bandwidth D_{bc} between ϕ_b and ϕ_c, and the bandwidth D_{cd} between ϕ_c and ϕ_d can be roughly calculated using Eq. (3) and Eq. (4), where F_s is the sampling frequency of speech signal.

$$D_{bc} = (\phi_c - \phi_b)/2\pi \times F_s \quad (3)$$

$$D_{cd} = (\phi_d - \phi_c)/2\pi \times F_s. \quad (4)$$

Watermarks are embedded by controlling the shape and power distribution between the last two formants. To minimize the distortions, formants are manipulated by shifting only one LSF, i.e., the penultimate LSF (last LSF but one), ϕ_c. If watermark $w = 0$, ϕ_c is shifted so that the bandwidth relationship between D_{bc} and D_{cd} is fixed as Eq. (5); if watermark $w = 1$, the bandwidth relationship is fixed as Eq. (6), where γ (γ>1.0) is used to control how much the formant is manipulated.

$$D_{bc} = \gamma \times D_{cd}, \quad w = 0 \quad (5)$$

$$D_{bc} = 1/\gamma \times D_{cd}, \quad w = 1. \quad (6)$$

Extration concept: As different bandwidth relationships have been established after embedding, watermarks w can be easily extracted by examining the relationships using Eq. (7).

$$w = \begin{cases} 0, & D_{bc} > D_{cd} \\ 1, & D_{bc} \leq D_{cd} \end{cases} \quad (7)$$

3. IMPLEMENTATION OF WATERMARKING

Figure 5(a) shows the watermarking embedding scheme. (i) Original signal, $x(n)$, is first segmented into frames, $x_m(n)$. STFT is applied to each frame to calculate its spectrogram. (ii) The sparse component of spectrogram is extracted using RPCA . (iii) LP analysis is applied to the sparse component to calculate the formants and LSFs. (iv) Each frame will be embedded with one-bit watermark by shifting the penultimate LSF according to Eq. (5) or Eq. (6). (v) LSFs will be converted back to LP coefficients to calculate the modified sparse component. (vi) The modified sparse component and low-rank component (obtained in (ii)) will be combined together and converted to

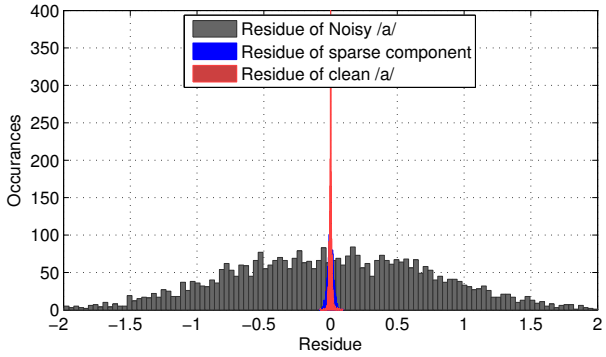

Fig. 3. Residue analysis for noisy speech.

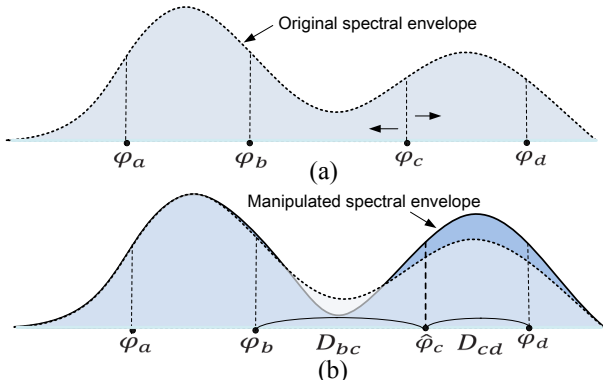

Fig. 4. Watermarking concept.

watermarked frame, $y_m(n)$, using inverse STFT (ISTFT). (vii) Finally, all watermarked frames will be connected together to construct the watermarked signal, $y(n)$. Figure 5(b) illustrates the watermark extraction scheme. The first three steps are the same as those in embedding scheme. When the LSFs are obtained from each frame, the bandwidths of formants can be calculated. Watermarks can be extracted using Eq. (7).

4. EVALUATIONS

The inaudibility and robustness of proposed method were evaluated with respect to embedding capacities (the proposed method is a blind method). All 12 speech in the ATR database (B set) (Japanese sentences: 8.1-sec, 20 kHz, and 16 bits) were used as stimuli [22]. λ_0 was fixed at 0.2 to attain the best result and γ was adopted as 1.75 to balance inaudibility and robustness. Embedded watermark was a random binary sequence. Embedding capacities were set as 4, 8, 16, 32, 64, 128, 200, and 400 bps. Evaluations were also done for three typical methods, i.e., LSB [23], DSS [24], CD [4, 25], and our previous method [14], which have separately exhibited good performance in inaudibility, robustness, and both inaudibility and robustness. Embedding capacities for these methods were 4, 8, 16, 32, 64, 128, and 256 bps according to their original implementations.

Inaudibility: Inaudibility was checked by log-spectrum distortion (LSD) [26] and perceptual evaluation of speech quality (PESQ) [27]. LSD in decibel (dB) measured the spectra distance between original signal and watermarked signal. LSD of 1.0 dB was chosen as the criterion and lower value indicated less distortion. PESQ in objective difference grades (ODGs) ranged from 0.5 (very annoying) to 4.5 (imperceptible) was used to evaluate the subjective quality, 3.0 (slightly annoying) was set as the criterion and higher value

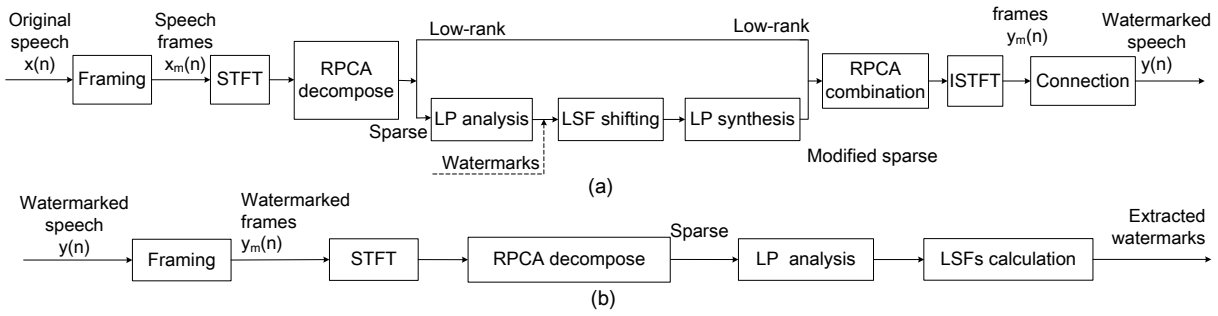

Fig. 5. Scheme for proposed watermarking: (a) embedding scheme and (b) detection scheme.

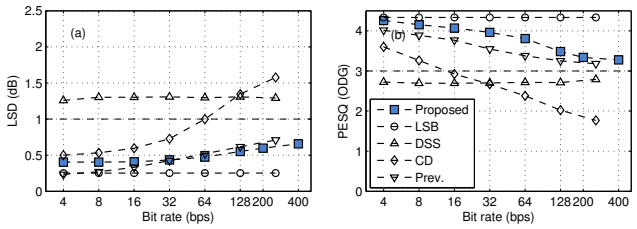

Fig. 6. Inaudibility of five methods.

indicated better quality. Evaluation results are plotted in Fig. 6. LSB had the best performance among the four methods. CD could satisfy inaudibility when embedding capacity was lower than 16 bps. DSS could not satisfy the criteria for either LSD or PESQ. Our previous method could satisfy the criteria for both LSD and PESQ. Nevertheless, the proposed method was better than CD, DSS, and our previous method, and it could satisfy inaudibility.

Robustness: Robustness was checked by Bit Detection Rate (BDR), i.e., the ratio between correctly extracted watermarks and embedded watermarks. BDR of 90% was set as the criterion and higher BDR indicated stronger robustness. Robustness was first evaluated against normal watermark extraction and speech processing including re-sampling at 24 kHz and 12 kHz, re-quantization with 24 bits and 8 bits, speech analysis/synthesis by gammatone filter-bank (GTFB) and STFT, and signal amplifying by 2.0 times. The BDR results are plotted in Fig. 7. It is clear that DSS performed the best. LSB was only robust against a few kinds of these processing. CD was robust against all processing except for re-quantization with 8 bits and GTFB. The proposed method was basically robust against all processing except for re-quantization with 8 bits. It was also better than our previous method at high embedding capacity.

We also applied four typical speech codecs to the watermarked speech, i.e., G.711, G723.1, G.726, and G.729. Figure 8 plots the BDR results. LSB was not robust against any speech codec; CD was only robust against G.711; DSS was not robust against G.723.1 and G.729. In contrast, the proposed method could survive from all speech codecs at low embedding capacity and its robustness against G.723.1 and G.729 was much improved than our previous method.

Discussion: This section compared the proposed method with other four methods. The proposed method could reach a balance in inaudibility and robustness, and its robustness was better than the other methods. A detailed discussion on how and why LSB, DSS, and CD performed can be found in [28]. The proposed method also outperformed our previous method in both inaudibility and robustness. Compared with the previous method, the scheme for formant manipulations was much simpler, indicating the performance of watermark extraction was more stable. Moreover, the introducing of RPCA enabled formant estimation to be more accurate which significantly improve the robustness of the proposed method.

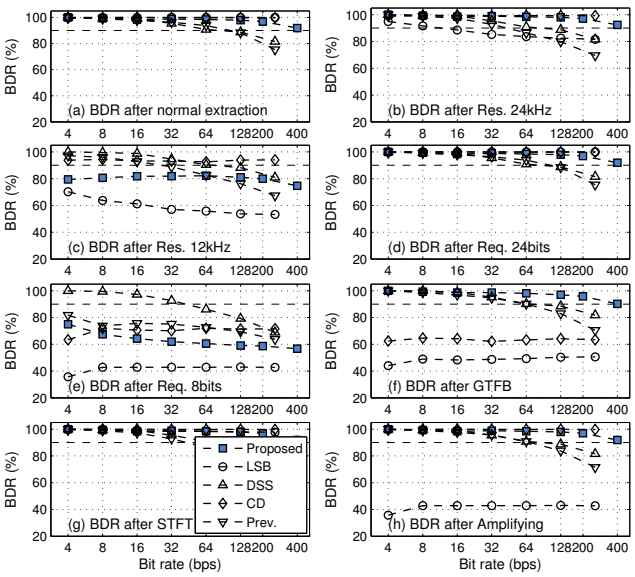

Fig. 7. Robustness of five methods against processing.

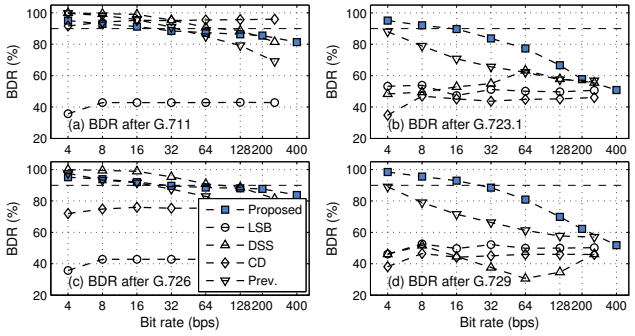

Fig. 8. Robustness of five methods against speech codecs.

5. CONCLUSIONS

This paper proposes a watermarking method for speech signals based on RPCA and formant manipulations. As the core information of speech tends to have a sparse structure and the noise/interferences have a low-rank structure, they are separable in T-F domain. RPCA is applied in the watermarking to extract the core information so that formants can be accurately estimated for watermark extraction under speech processing/codecs, which improves the robustness of proposed method. LP analysis is used to extract formants and watermarks are embedded into formants by controlling LSFs. Benefit from these considerations, the proposed method exhibites better performance in inaudibility and robustness than the other methods.

6. REFERENCES

[1] B. C. J. Moore, "An Introduction to the Psychology of Hearing," *6th Edition, Brill Academic Pub.*, 2013.

[2] H. Guang , G. Jonathan , and V. L. L. Thing, "Time-spread echo-based audio watermarking with optimized imperceptibility and robustness," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 227-239, 2015.

[3] K. Hofbauer, G. Kubin, and W. Bastiaan Kleijn, "Speech watermarking for analog flat-fading bandpass channels," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 8, pp. 1624–1637, 2009.

[4] M. Unoki and D. Hamada, "Method of digital-audio watermarking based on cochlear delay characteristics," *J. Inn. Com. Inf., and Cont.*, vol. 6, no.(3(B)), pp. 1325–1346, 2010.

[5] B. Chen and G. W. Wornel, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Information Theory*, vol. 47, no. 4, pp. 1423–1443, 2001.

[6] N. M. Ngo, B. M. Kurkoski, and M. Unoki,, "Robust and reliable audio watermarking based on dynamic phase coding and error control coding," *Proc. EUSIPCO*, pp. 2316–2320, 2015.

[7] H. S. Malvar and A. F. Florłncio, "Improved spread spectrum: A new modulation technique for robust watermarking," *IEEE Trans. Signal Processing*, vol. 51, no. 4, pp. 898–905, 2003.

[8] D. Kirovski and H. Malvar, "Robust spread spectrum audio watermarking," *Proc. ICASSP*, vol. 3, pp. 1345–1348, 2001.

[9] K. Reza, F. Prez-Gonzlez, A. A.Mohammad, and B. Fereydoon , "Data hiding robust to mobile communication vocoders," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 2345–2357, 2016.

[10] E. Chrysochos, V. Fotopoulos, M. Xenos, and A. N. Skodras, "Hybrid watermarking based on chaos and histogram modification," *Journal of Signal, Image and Video Processing*, vol. 8, no. 5, pp. 843–857, 2014.

[11] P. W. Chan, M. R. Lyu, and R. T. Chin, "A novel scheme for hybrid digital video watermarking: approach, evaluation and experimentation," *IEEE Trans. Circuit and system for video technology*, vol. 15, no. 12, pp. 1638–1649, 2005.

[12] B. Y. Lei, K. T. Lo, and H. j. Lei, "Hybrid SVD-based audio watermarking scheme," *Proc. Communications, Circuits and Systems (ICCCAS)*, pp. 428–432, 2010.

[13] S. Wang and M. Unoki, "Watermarking method for speech signals based on modifications to LSFs," *Proc. IIHMSP*, pp. 283–286, 2013.

[14] S. Wang and M. Unoki, "Watermarking of speech signals based on formant enhancement," *Proc.EUSIPCO*, pp. 1257–1261, 2014.

[15] P. S. Huang, S. D. Chen, P. Smaragdis, M. H. Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," *Proc. ICASSP*, pp. 57– 60, 2012.

[16] E. J. Candes, X. D. Li, Y. Ma, and J. Wright, "Robust principal component analysis," *J. ACM*, vol. 58, pp. 1111–1137, 2011.

[17] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[18] B. S. Atal, "The history of linear prediction," *IEEE Signal Processing Magazine*, vol. 23, pp. 154–161, 2006.

[19] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *Journal of the Acoustical Society of America*, vol. 57. no. 537(A), pp. 35–35, 1975.

[20] M. H. Johnson, "Line spectral frequencies are the poles and zeros of a discrete matched-impedance vocal tract model," *Journal of the Acoustical Society of America*, vol. 108, no. 1, pp. 457-465, 2000.

[21] T. Bckstrm, P. Alku, T. Paatero, and B. Kleijn, "A time-domain interpretation for the LSP decomposition," *IEEE Trans. Speech and Audio Processing*, vol. 12, pp. 554–560, 2004.

[22] K. Takeda et al, "Speech database user's manual," *ATR Technical Report* TR-I-0028, 2010.

[23] P. Bassia and I. P. Pitas, "Robust audio watermarking in the time domain," *Proc. EUSIPCO*, pp. 25–28, 1998.

[24] L. Boney, H. H. Tewfik, and K. H. Hamdy, "Digital watermarks for audio signals," *Proc. ICMCS*, pp. 473-480, 1996

[25] M. Unoki and R. Miyauchi, "Reversible watermarking for digital audio based on cochlear delay characteristics," *Proc. IIHMSP*, pp. 314–317, 2011.

[26] A. Gray, Jr., and J. Markel, "Distance measures for speech processing," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.

[27] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[28] S. Wang and M.Unoki, "Speech watermarking method based on formant tuning," *IEICE Trans. INF. & SYST., Enriched Multimedia*, vol. E98-D, no. 1, pp. 29–37, 2015.