Speech Information Hiding by Modification of LSF Quantization Index in CELP Codec

Candy Olivia Mawalim*, Shengbei Wang†, Masashi Unoki*
* Japan Advanced Institute of Science and Technology, Japan
E-mail: {candyolivia,unoki}@jaist.ac.jp
† Tianjin Polytechnical University, China
E-mail: wangshengbei@tiangong.edu.cn

***Abstract*—A prospective method for securing digital speech communication is by hiding the information within the speech. Most of the speech information hiding methods proposed in prior research are lacking in robustness when dealing with the encoding process (e.g. the code-excited linear prediction (CELP) codec). The CELP codecs provide a codebook that represents the encoded signal at a lower bit rate. As essential features in speech coding, line spectral frequencies (LSFs) are generally included in the codebook. Consequently, LSFs are considered as a prospective medium for information hiding that is robust against CELP codecs. In this paper, we propose a speech information hiding method that modifies the least significant bit of the LSF quantization obtained by a CELP codec. We investigated the feasibility of our proposed method by objective evaluation in terms of detection accuracy and inaudibility. The evaluation results confirmed the reliability of our proposed method with some further potential improvement (multiple embedding and varying segmentation lengths). The results also showed that our proposed method is robust against several signal processing operations, such as resampling, adding Gaussian noise, and several CELP codecs (i.e., the Federation Standard-1016 CELP, G.711, and G.726).**

I. Introduction

Technological developments supporting the speech communication systems via the public switched telephone network (PSTN) and Voice over Internet Protocol (VoIP) have advanced significantly over recent years. However, the use of such advanced technology raises security concerns, for example, risks of tampering due to an insecure transmission over communication channels. Ensuring the security of a speech communication system is crucial not only to protect the privacy and confidentiality of military or government-related communication but also for our daily mobile and voice communication via the Internet [1]. Consequently, techniques to secure speech and voice communication systems have attracted a great deal of attention among researchers, particularly in the field of speech information hiding.

Several techniques have been proposed for hiding information in speech signals; for example, the least significant bit (LSB), phase modulation, and direct spread spectrum (DSS) [2]. However, each of the classical techniques has shortcomings, especially in controlling the trade-off between inaudibility and robustness. Recent methods have enhanced those techniques by applying concepts in psychoacoustics to deal with this trade-off. For instance, cochlear delay-based information hiding was proposed to improve the phase modulation technique by utilizing the delay characteristics in the baseline membrane of the human cochlear [3]. Although the improved techniques could successfully outperform the classical methods, the inaudibility-robustness trade-off issue remains inextricable for speech information hiding methods, especially when dealing with the speech codec applied to a digital communication system [1], [4], [5].

A speech codec encodes and decodes speech signals into digital information before storing or transmitting them through a communication channel. The most advanced and widely used coding algorithm comes from the code-excited linear prediction (CELP) family [6]. The CELP codec can provide high-quality speech at a low bitrate [6], [7]. To ensure the robustness of a speech information hiding method for speech codecs, such as the CELP codec, in-encoder or analysis-by-synthesis (AbS) approaches have been considered [4]. For example, several prior studies proposed speech information hiding methods based on a specific speech codec [1], [8] or using the AbS approach based on linear predictive coding parameters [5], [9], [10], [11].

Line spectral frequencies (LSFs) are one of the parameters derived by linear predictive coding that is commonly used in speech technology, including information hiding. It provides strong robustness for information hiding in dealing with speech coding algorithms compared with other typical methods [1], [5], [9], [10]. For example, a direct modification of LSFs for a speech watermarking method using dither modulation-quantization index modulation (DM-QIM) was proposed in [10]. Unfortunately, this method is weak against several signal processing operations. To improve the robustness, Wang et al. proposed an LSFs modification-based speech watermarking technique based on the concept of formant tuning [5], [9]. A linear prediction analysis was conducted to estimate the formants of speech signal in each frame. Subsequently, the formant tuning was performed by controlling the formant bandwidth with regards to the desired watermark bit.

In this research, we propose an LSF modification-based speech information hiding technique that considers the quantization process in a specific speech codec (the Federal Standard-1016 (FS-1016) CELP codec [12]). Although more advanced CELP codecs have been proposed recently, the

simplicity of the FS-1016 CELP codec can provide a general framework since the core element of AbS is clearly represented. Moreover, perceptual criteria and good interpolation properties were also considered in its development. We expect that our proposed method could provide a strong robustness for speech codecs (LSF modification based on a specific codec) and an inaudible watermark (LSB and perceptual properties are preserved in the FS-1016 codec). Our experiments aim to investigate the feasibility of the hiding process by modifiying the LSF quantization index in the CELP codec. Moreover, we also investigate the flexibility of our proposed method with two different speech datasets and analysis parameters. Finally, a comparative evaluation with two typical methods, log-spectrum distortion (LSD) and DSS, is conducted under normal and several signal processing attacks conditions to investigate the robustness of our proposed method.

II. Background Concept

In this section, the key concepts underlying our proposed method are described. We begin by introducing the CELP codecs. We then explain the concept of LSFs followed by the LSF quantization process based on the FS-1016 CELP codec.

A. CELP Codecs

CELP codecs are the most common speech codecs used in digital communication systems due to their low-bitrate high-quality speech representation [6], [7]. These codecs are based on the source-system model that mimics the human speech production mechanism [6], [7], [13]. Atal and Schroeder proposed a typical CELP codec based on AbS linear predictive coding [7]. Linear predictive coding attempts to estimate vocal tract parameters by estimating a current speech signal using a linear combination of past samples. The following differential equation characterizes the mathematical form of linear predictive coding:

$$s(n) = \sum_{i=1}^{M} a(i)s(n-i) + e(n) \tag{1}$$

where $a(i)$ corresponds to the filter coefficient in i-th order, M is the maximum order of the prediction (typically 10), and $e(n)$ is the prediction error.

The transfer function for the corresponding linear prediction differential equation is represented by tenth-order all-pole autoregressive filters, which is given by:

$$H(z) = \frac{1}{1 - \sum_{i=1}^{10} a(i)z^{-i}} \tag{2}$$

Figure 1 illustrates the source-system model by AbS linear prediction. In CELP coding, the excitation generator generates an excitation vector codebook $\mathbf{x}$ by minimizing the residual error e, which can be written mathematically as,

$$e^{(i)} = s_w - \hat{s}_w^0 - g^{(i)}\hat{s}_w^{(i)} \tag{3}$$

where s_w is a vector of perceptually-weighted input speech, $\hat{s}_w^0$ is the initial filter state output vector, $g^{(i)}$ is the gain factor,

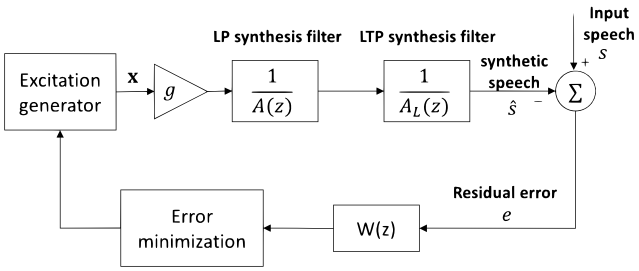

Fig. 1. Source-system model in AbS linear prediction.

and $\hat{s}_w^{(i)}$ is the synthetic speech vector associated with the $\mathbf{x}^{(i)}$ with i as the codebook index.

In standard AbS linear prediction algorithms, the tenth-order short term linear prediction is used as the linear prediction synthesis filter ($1/A(z)$). $A(z)$ denotes the line spectrum pairs (LSPs) that can be given by:

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + ... + a_{10} z^{-10} \tag{4}$$

where a_i is the i-th order linear prediction coefficients (LPCs).

The long term prediction (LTP) synthesis filter ($1/A_L(z)$) captures the long-term correlation and represents the speech periodicity mechanism. The perceptual weighting filter $W(z)$ models errors by masking the quantization noise with high-energy formants. The perceptual weighting filter $W(z)$ can be written as follows:

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} = \frac{1 - \sum_{i=1}^{M} \gamma_1(i)a(i)z^{-i}}{1 - \sum_{i=1}^{m} \gamma_2(i)a(i)z^{-i}} \tag{5}$$

where γ_1 and γ_2 are the adaptive weights that satisfy $0 < \gamma_1 < \gamma_2 < 1$, and m is the order of the linear predictor. γ_1 ranges between 0.94 and 0.98 and γ_2 ranges between 0.4 and 0.7 depending on the tilt or flatness characteristics of the linear prediction spectral envelope [13], [14].

B. LSFs Concept

Direct quantization of LPCs, $a(i)$, is commonly not applicable in standardized coding algorithms due to its sensitivity. A slight modification to LPCs can cause a significant distortion in the speech since it raises loss to the filter stability. In other words, directly altering the LPCs will most likely causes the poles to be positioned outside the unit circle. Due to this reason, another quantization method is preferable. In the CELP-based speech coding algorithm, LSPs are generated due to their superior quantization characteristics [15], [16].

As described in Eq. (4), the LSPs are typically a tenth-order polynomial. This polynomial is computed using two auxiliary polynomials $P(z)$ and $Q(z)$, which are given by:

$$P(z) = A(z) + z^{-11}A(z^{-1}) \tag{6}$$

$$Q(z) = A(z) - z^{-11}A(z^{-1}) \tag{7}$$

where $P(z)$ is a symmetric polynomial, and $Q(z)$ is an anti-symmetric polynomial. $P(z)$ and $Q(z)$ consist of five complex conjugate pairs of zeros that typically lie on the unit circle.

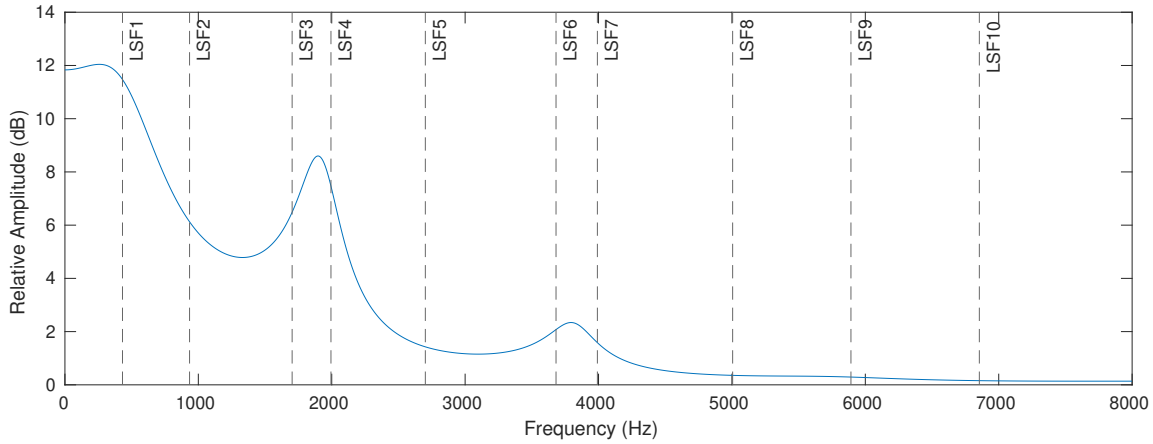

Fig. 2. Example of the frequency response of a linear predictive filter overlaid with the corresponding LSFs obtained from the tenth-order linear predictive analysis of a 25-ms-long voiced speech segment.

TABLE I
LSF QUANTIZATION MATRIX IN FS-1016 CELP CODEC

		Quantization Index															
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
LSF Index	1	100	170	225	250	280	340	420	500	0	0	0	0	0	0	0	0
	2	210	235	265	295	325	360	400	440	480	520	560	610	670	740	810	880
	3	420	460	500	540	585	640	705	775	850	950	1050	1150	1250	1350	1450	1550
	4	620	660	720	795	880	970	1080	1170	1270	1370	1470	1570	1670	1770	1870	1970
	5	1000	1050	1130	1210	1285	1350	1430	1510	1590	1670	1750	1850	1950	2050	2150	2250
	6	1470	1570	1690	1830	2000	2200	2400	2600	0	0	0	0	0	0	0	0
	7	1800	1880	1960	2100	2300	2480	2700	2900	0	0	0	0	0	0	0	0
	8	2225	2400	2525	2650	2800	2950	3150	3350	0	0	0	0	0	0	0	0
	9	2760	2880	3000	3100	3200	3310	3430	3550	0	0	0	0	0	0	0	0
	10	3190	3270	3350	3420	3490	3590	3710	3830	0	0	0	0	0	0	0	0

These two polynomials can be regarded as an interconnected tube representation of the vocal tract in a speech production system [16]. The linear combination of these two polynomials represents the actual resonance $A(z)$, which is given by:

$$A(z) = \frac{P(z) + Q(z)}{2} \tag{8}$$

The roots of the two polynomials $P(z)$ and $Q(z)$ are referred to as LSFs, which are associated with speech formants [16]. The relationship between LSFs and the frequency response of a linear prediction filter is shown in Fig. 2. Speech formants are important aspects of speech perception. Due to this fact, the importance level of formants is considered in the quantization process in the coding algorithm [16]. For example, on the basis of the example in Fig. 2, the LSF lines 5 and 6 may be related to the formant F2. However, since the formant F2 is less important than formant F1 (represented by lines 3 and 4), the quantization representation in CELP codecs for the LSF lines 3 and 4 is more detailed than LSF lines 5 and 6.

C. FS-1016 CELP Quantization Algorithm for LSFs

In standardized CELP codecs, three to four bits are allocated as quantization bits to represent each LSF extracted from Eqs. 6 and 7. In this work, we utilized the FS-1016 CELP quantization algorithm as the study case for quantizing the LSFs. The FS-1016 CELP codec is one of the first-generation CELP codecs that operates at a bitrate of 4.8 kb/s. This standard configuration is based on gain-shape vector quantization and is designed for 8-kHz sampled speech segmented into 30-ms intervals. We chose this algorithm because the core element of the CELP codec (AbS linear prediction coding) is clearly represented, and thus, adapting the current technique to other more advanced algorithms is possible. Furthermore, the simplicity of the FS-1016 CELP quantization process is derived from using perceptual criteria and good interpolation properties [12], [13]. This criteria is based on the special properties when the LSPs $A(z)$ is in the minimum phase condition. In the minimum phase condition, the zeros lie on the unit circle, and the zeros of the two polynomials are interlaced [15]. These properties are perceptually meaningful, which should be preserved after quantization [17].

The FS-1016 CELP quantization algorithm uses an independent, non-uniform scalar quantization procedure. The quantization of LSFs is based on the quantization matrix (as shown in Table. I). There are 34 bits per frame that represent the LSFs. Three bits are used for representing LSF 1 and LSFs 6 to 10. Four bits are allocated for representing

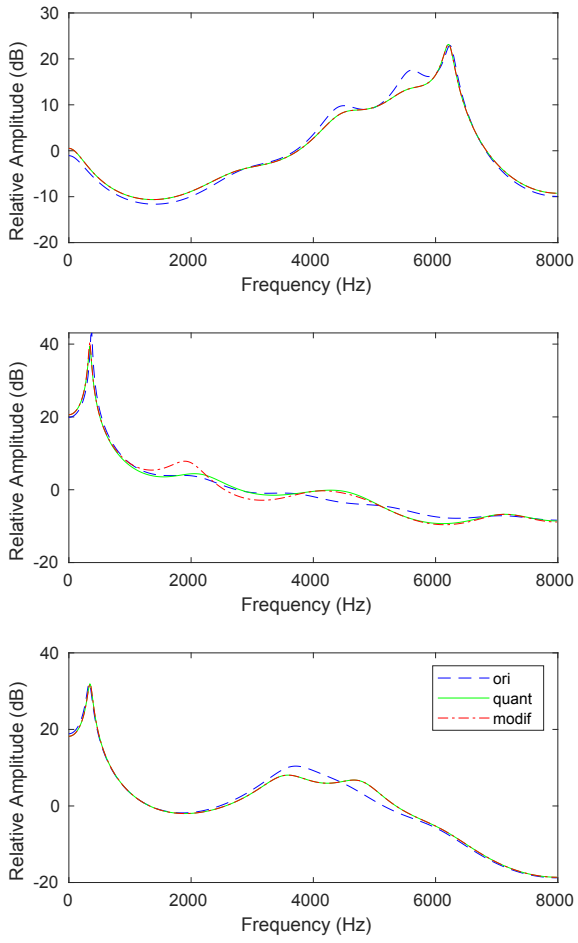

Fig. 3. Frequency response spectra from actual LSFs (ori), quantized LSFs (quant), and modification of least significant quantized LSFs (modif).

(LSFs 2 to 5). The quantization procedure may result in non-monotonicity, which leads to the loss of the minimum phase condition (ill-conditioned case) [18]. Accordingly, after the quantization process, an adjustment process is required to restore the monotonicity.

III. Speech Information Hiding Scheme

We propose a speech information hiding method by modifying the least significant LSF quantization bit. In this section, we explain the LSF quantization bit modification, embedding, and detection process in detail.

A. LSFs Quantization Bits Modification

Instead of direct quantization of LSFs, we follow the FS-1016 gain-shape vector quantization to preserve the robustness of the information hiding method for this specific speech codec. The adjustment process in the FS-1016 CELP codec, as mentioned in Subsection II-C, applies a slight modification to the LSB of the allocated bits for LSFs. Since the speech distortion caused by this adjustment is minor, it is promising to obtain an inaudible speech information hiding method by modifying the most insignificant bit of the allocated LSF quantization index.

Figure 3 shows the impact in frequency response spectra changing caused by the LSF modification on the basis of the FS-1016 CELP quantization algorithm. From this figure, we can see that the frequency response spectra are shifted when the quantization process is performed. Despite this shifting, the impact is less significant because this quantization algorithm is based on perceptual criteria properties (e.g., the higher formant is less meaningful in perception). Moreover, this figure also shows that since the embedding is based on the standardized CELP quantization method, the different spectra between the quantized LSFs and modified LSFs are insignificant (potential for inaudible modification).

B. Embedding

Figure 4 (top) shows the embedding process of our proposed method. There are five main steps as follows:

1) The input speech $s(n)$ is segmented into non-overlapping t-length-frames. t denotes the time length in ms (which we will use as our independent variable in Section IV).
2) A 10-th order linear prediction (LP) filter is used to analyze the framed input signal to obtain the 10 LPCs $a(i)$, where $i = 1, 2, ..., 10$.
3) The LPCs $a(i)$ obtained from the previous step are converted to LSF quantization bits on the basis of the FS-1016 CELP quantization mechanism by using the following substeps:
   a) generating the LSP polynomials $P(z)$ and $Q(z)$ on the basis of Eqs. (6) and (7) with regard to the LPCs $a(i)$;
   b) computing both zeros from symmetrical and anti-symmetrical polynomials on the basis of Descartes' rule to obtain the LSFs;
   c) quantizing the LSFs on the basis of the LSF quantization matrix in Table. I to obtain the LSF quantization indexes;
   d) adjusting the LSF quantization indexes to preserve monotonicity by checking and correcting the ill-conditioned cases;
   e) converting the adjusted LSF quantization indexes to a binary form as LSF quantization bits.
4) The least significant LSF quantization bits are manipulated in accordance with the watermark bit stream w. After the modification, the dequantization process is performed to obtain the modified LSP coefficients $p'(i)$ and $q'(i)$. Next, these coefficients are converted to LPCs $a'(i)$.
5) Finally, the watermarked speech $s'(n)$ is obtained by

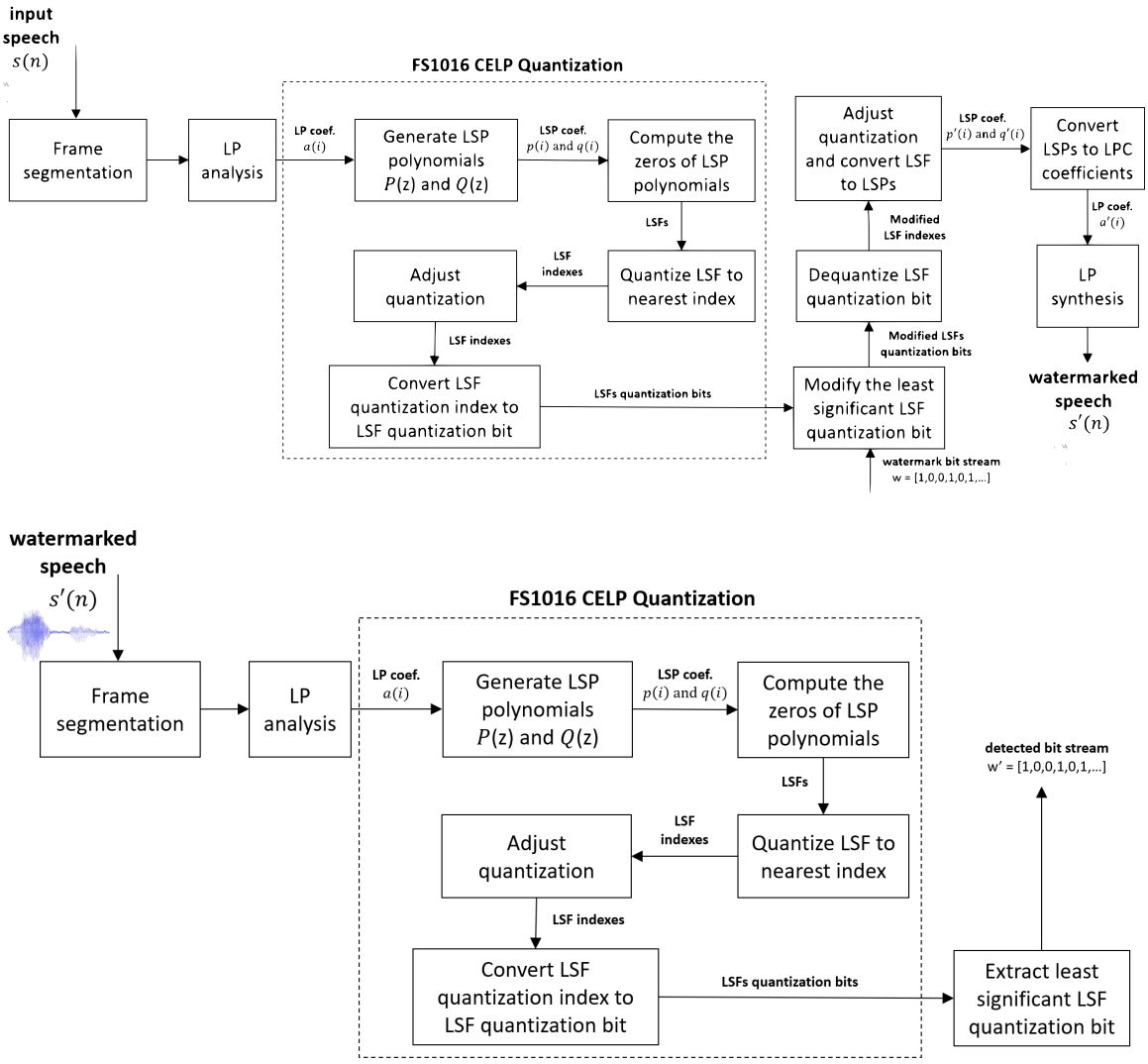

Fig. 4. Block diagram of proposed speech information hiding: (top) embedding process and (bottom) detection process.

using LP synthesis in accordance with the modified LPCs $a'(i)$.

C. Detection

Figure 4 (bottom) shows the detection process of our proposed method. It begins with the first three steps of our proposed embedding process with the watermarked signal $s'(n)$ as the input. Subsequently, we extract the least significant LSF quantization bit as the detected watermarks w'.

IV. Evaluation and Discussion

We evaluated our proposed method using several scenarios to check the feasibility and robustness of our proposed method. First, we investigated our method's feasibility by using the designated configuration (input signals and analysis parameters) of the FS-1016 CELP codec. Then, we utilized another speech dataset with a different configuration to investigate our method's flexibility despite the various input and analysis parameters. We also investigated the possibility of enhancing the robustness and payload of our method. Finally, we compared our method with a typical speech information hiding method, such as LSB and DSS [2], under normal and several signal processing attacks conditions.

A. Evaluation Criteria

We performed an objective evaluation to measure the robustness and inaudibility of our proposed method. We calculated the bit error rate (BER) in $\%$ for the robustness evaluation, and calculated the LSD [19] and perceptual evaluation of speech quality (PESQ) [20] ITU-T P.862 for the inaudibility evaluation. The BER determines the detection accuracy (the number of incorrectly detected watermark bits over the total

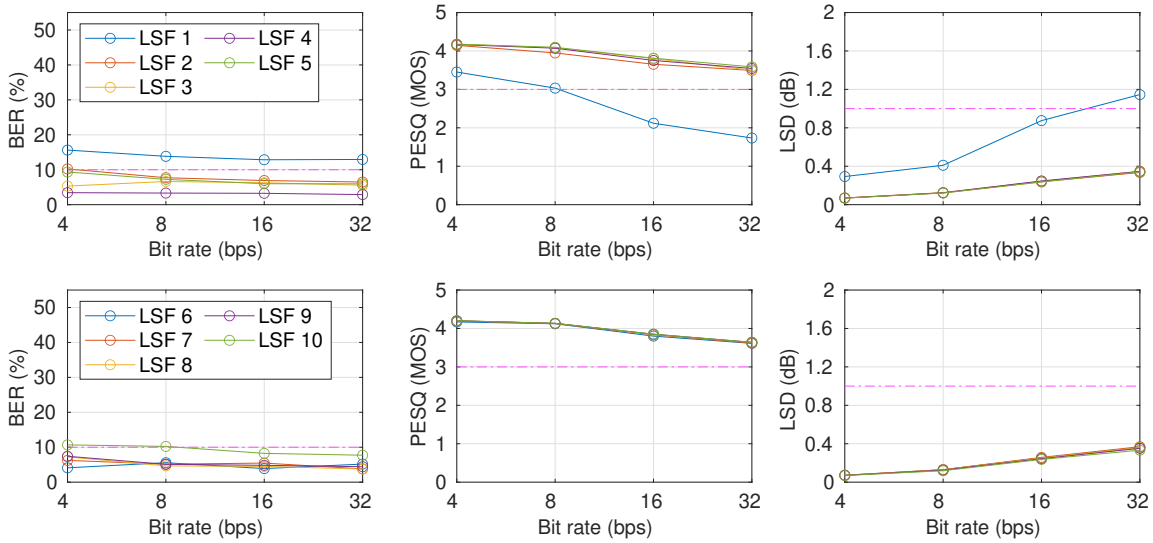

Fig. 5. Objective evaluation of our proposed method in each LSF quantization bit by using BER, PESQ, and LSD in the original FS-1016 CELP quantization algorithm configuration. The input signal is sampled at 8 kHz and its frame segmentation length t is 30 ms.

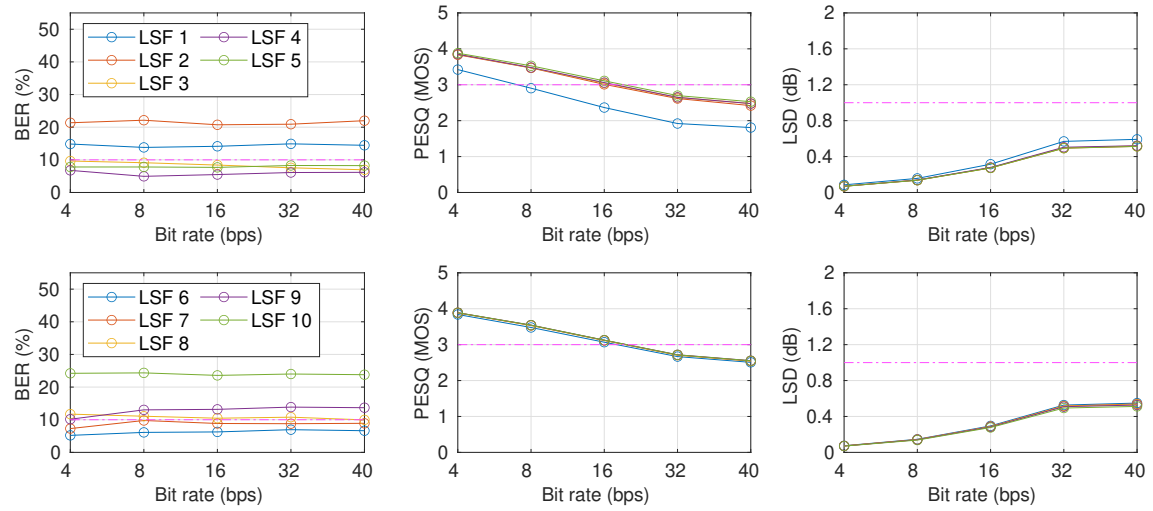

Fig. 6. Objective evaluation of our proposed method in each LSF quantization bit by using BER, PESQ, and LSD in the adapted quantization configuration. The input signal is sampled at 16 kHz and its frame segmentation length t is 25 ms.

number of embedded watermark bits). The LSD determines the spectral distortion of the watermarked signal in comparison with the original signal in decibels (dB). In information hiding, the typical threshold for LSD is 1 dB. The PESQ determines the perceptual speech quality, which models mean opinion scores (MOSs) that vary from a scale of 1 (bad) to 5 (excellent). The typical threshold for PESQ in information hiding is 3 (fair, slightly annoying).

B. Basic Evaluation

The basic evaluation follows our first evaluation scenario. This evaluation aims to check the feasibility of hiding information in the least significant LSF quantization indexes using the FS-1016 CELP codec. As per the aforementioned description of the FS-1016 CELP algorithm, an opensource dataset (VoxForge) with ten selected English-spoken speech stimuli was used in the first evaluation scenario. Each stimulus in this dataset is sampled at 8 kHz with 16-bit quantization. The duration of each stimulus ranges between five and ten

seconds. The frame length parameter is 30 ms, which is the same as that in the coding algorithm. Since a fixed frame length and one LSF channel is used, the maximum available payload is only 33 bps. Due to this limitation, we analyzed the performance of our proposed method in various bitrates (4, 8, 16, and 32 bps).

Figure 5 shows the result of the basic evaluation. This figure confirmed the feasibility of hiding information in the speech by the proposed method. The adequate detection rate could be obtained despite the watermark position in any LSF, except LSF 1. The modification of LSF 1 caused a significant distortion to the watermarked signal. LSF 1 often represents the first formant that is significantly meaningful for speech perception. Thus, changing this parameter is not recommended for information hiding.

The inaudibility of our proposed method can be represented in Fig. 5 at the second and the third columns. The perceptual quality of the watermarked signal is good (PESQ score almost around four4, even in the at a high-bitrate). Along with the perceptual quality, the sound distortion is also small enough (LSD is less than 1 dB).

C. Robustness Evaluation

Unlike the input parameter in the FS-1016 CELP codec, we utilized the ATR Japanese speech dataset (B set) [21], which is sampled at 16 kHz, to investigate our proposed method's robustness. Twelve stimuli were selected from this dataset for our evaluation. Each signal in this dataset has an 8.1-sec duration length. In this subsection, we aim to investigate whether our method can work regardless of the different input and analysis parameters.

Figure 6 shows the objective evaluation results of our proposed method with a 25-ms-long analysis-synthesis frame. Although there is a slight drop in performance, the overall result in this scenario ties well with that shown in Fig. 5. In most cases, the robustness and inaudibility when hiding in each LSF are sufficient (BER around 10%, PESQ around 3, and LSD less than 1 dB), except for LSFs 1, 2, and 10. Thus, LSFs 1, 2, and 10 are not recommended as embedding mediums.

In summary, this result highlights that our proposed method is robust enough to deal with different segmentation lengths and input signals sampled at the different sampling frequencies. The compression in the quantization process does not cause significant defects in the embedded watermarks. Moreover, due to the fact that the process in our proposed method is based on AbS with the FS-1016 CELP codec, the robustness for this coding algorithm can be assured.

D. Further Potential Improvement

One of the straightforward methods to improve the robustness and payload of our proposed method is by using multiple embedding or reducing the duration of the analysis-synthesis frame. In this subsection, we investigate the ways to improve the robustness and payload considering the impact of degradation in sound quality.

TABLE II
EVALUATION RESULT FOR MULTIPLE EMBEDDING IN THREE SELECTED LSFS (LSF 4, 6, AND 7)

Variable	**Evaluation Score**	**bit rate (bps)**				
		12	**24**	**48**	**96**	**120**
Payload	BER %	8.54	8.59	8.7	8.56	8.54
	PESQ (MOS)	3.81	3.39	3.00	2.57	2.42
	LSD (dB)	0.08	0.16	0.33	0.59	0.61
		4	**8**	**16**	**32**	**40**
Robustness	BER (%)	3.73	3.73	3.36	3.31	3.73
	PESQ (MOS)	3.80	3.39	2.99	2.58	2.39
	LSD (dB)	0.08	0.16	0.32	0.58	0.60

1) Modification of Multiple LSFs:

On the basis of the detection accuracy evaluation results in Figs. 5 and 6, we selected the three most robust least significant LSF quantization bits (LSFs 4, 6, and 7). We checked the improvement of payload and robustness performance of multiple bit embedding in consideration of the sound quality degradation impact. The input signal and analysis parameter followed the evaluation in Subsection IV-C. The evaluation for payload improvement was conducted by inserting three different watermark binaries into the selected LSFs. Thus, the payload could be improved threefold. Another evaluation for robustness improvement followed the repetitive coding concept. A watermark bit is duplicated into three watermarks, which were then embedded into the selected LSFs. The detected watermark bit was determined by calculating the mean value of those three watermarks and classifying them into binary 0 or 1 with a threshold of 0.5.

Table II shows the evaluation results for multiple embedding. The results suggest that multiple embedding could improve both the payload and robustness of the proposed method with an almost similar sound quality with single embedding. By embedding three different watermark bits into three LSFs, the detection accuracy is also similar to that of single embedding (BER is less than 10%). This result shows that we can also attempt to embed the watermark stream into other LSF quantization bits (LSF 3 5 8, and 9) as a further prospective improvement. Moreover, the evaluation result of multiple embedding also shows that we could use repetitive coding if our system requires a higher detection accuracy (BER is less than 5%).

2) Varying the Frame Segmentation Length:

Improving the payload robustness is also likely to be achieved by reducing the fixed frame segmentation length of t. Figure 7 shows the comparison result of the objective evaluation using BER, PESQ, and LSD where the frame segmentation varies from 5 to 25 ms. This result indicates that the high detection accuracy could be achieved at a high bitrate, although the frame segmentation length is short (BER is less than 10%). As for the inaudibility evaluation result, our proposed method could satisfy the threshold of the LSD score even at a higher bitrate. In contrast, the result of the PESQ evaluation shows the constraint in speech quality degradation at a higher bitrate.

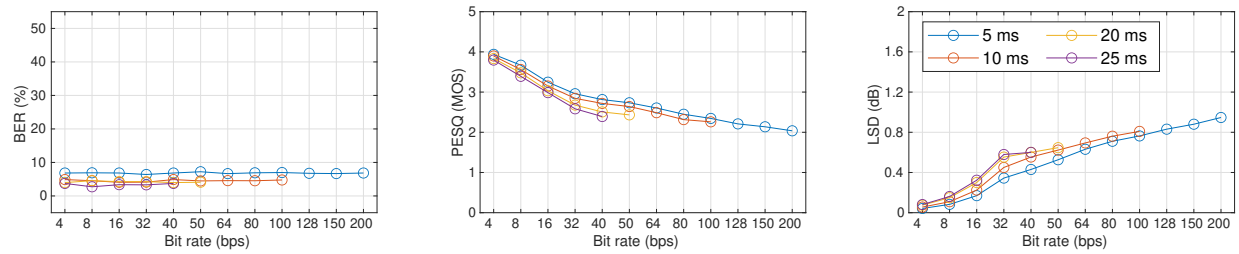

Fig. 7. Objective evaluation results of the proposed method in comparison with several frame segmentation lengths (5, 10, 20, and 25 ms).

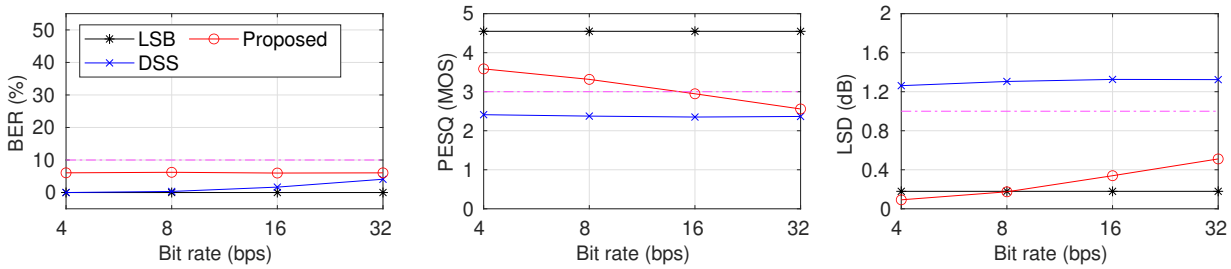

Fig. 8. Objective evaluation result of comparative methods under normal conditions based on detection accuracy (BER) and inaudibility (PESQ and LSD).

TABLE III
OPTIMIZATION RESULT USING A COMBINATION OF MULTIPLE EMBEDDING AND VARYING FRAME LENGTHS.

Bit rate (bps)	50	100	200	400	800	1600
BER(%)	4.033	5.144	6.626	10.761	11.461	18.816
LSD	0.144	0.292	0.313	0.636	0.901	1.065
PESQ	3.499	3.042	2.996	2.479	2.065	1.833

3) Optimizing Multiple Embedding and Variation of Frame Segmentation Length:

Improving payload and detection accuracy could be achieved using either multiple embedding or a shortened frame length (as shown in Table. II and Fig. 7). On the basis of these results, we optimized the embedding capacity by using both these methods. First, we improved the detection accuracy by assigning weights to each LSF on the basis of the detection accuracy obtain in the basic evaluation. Subsequently, on the basis of the weights, we performed majority voting to determine a detected watermark. Finally, we preserved the speech quality by not embedding to LSF 1 and 2, and optimizing the repetitive embedded bits (minimizing the embedded bits but preserving the accuracy) for a watermark.

E. Comparison with Typical Speech Information Hiding Methods

We performed a comparative evaluation between our proposed method (single embedding in LSF 4) and two typical methods (LSB and DSS) with the objective evaluation of robustness and inaudibility. The traditional LSB method alters the most insignificant quantization bits of the speech signal with watermarks to maintain the inaudibility of the distortion. In contrast, the DSS method spreads the desired watermarks over the whole frequency band to ensure robustness. The comparative evaluation was conducted by using the ATR dataset, as mentioned in Subsection IV-C. The bitrate ranges from 4 to 32 bps.

Figure 8 shows the comparative evaluation result under the normal condition (without considering any attacks). This result indicates that the LSB method could achieve a high accuracy and inaudibility even at a high bitrate. IN contrast, the DSS method caused a significant distortion to the watermarked signal despite the high detection accuracy. Our proposed approach works in between the LSB and DSS methods. Although it could not achieve perfect accuracy (BER is less than 10%), our proposed method could achieve better inaudibility compared with the DSS method. In other words, we could say that our proposed method is reliable (robust and inaudible) at low bitrates (up to 16 bps for single embedding).

In the actual speech communication system, several signal processing operations often invaded the transmitted speech. Figure 9 denotes the robustness evaluation result of our comparative methods against several signal processing operations. In most cases, the LSB method (blue line) is very fragile against any attacks, whereas the DSS method (orange line) is very robust. Even though it is not as robust as the DSS method, our proposed method (yellow line) could provide robustness against several operations. Figure 9(a) confirmed our hypothesis that our method is robust against the specific FS-1016 CELP codec. Moreover, our proposed method is also robust against noise addition (AWGN) (Fig. 9(b)), resampling (Fig. 9(c–d)), and requantization to higher bits (Fig. 9(f)). The requantization to lower bits (Fig. 9(e)) remains as a limitation robustness in our proposed method. However, our proposed method is somewhat robust against other speech codecs, e.g.

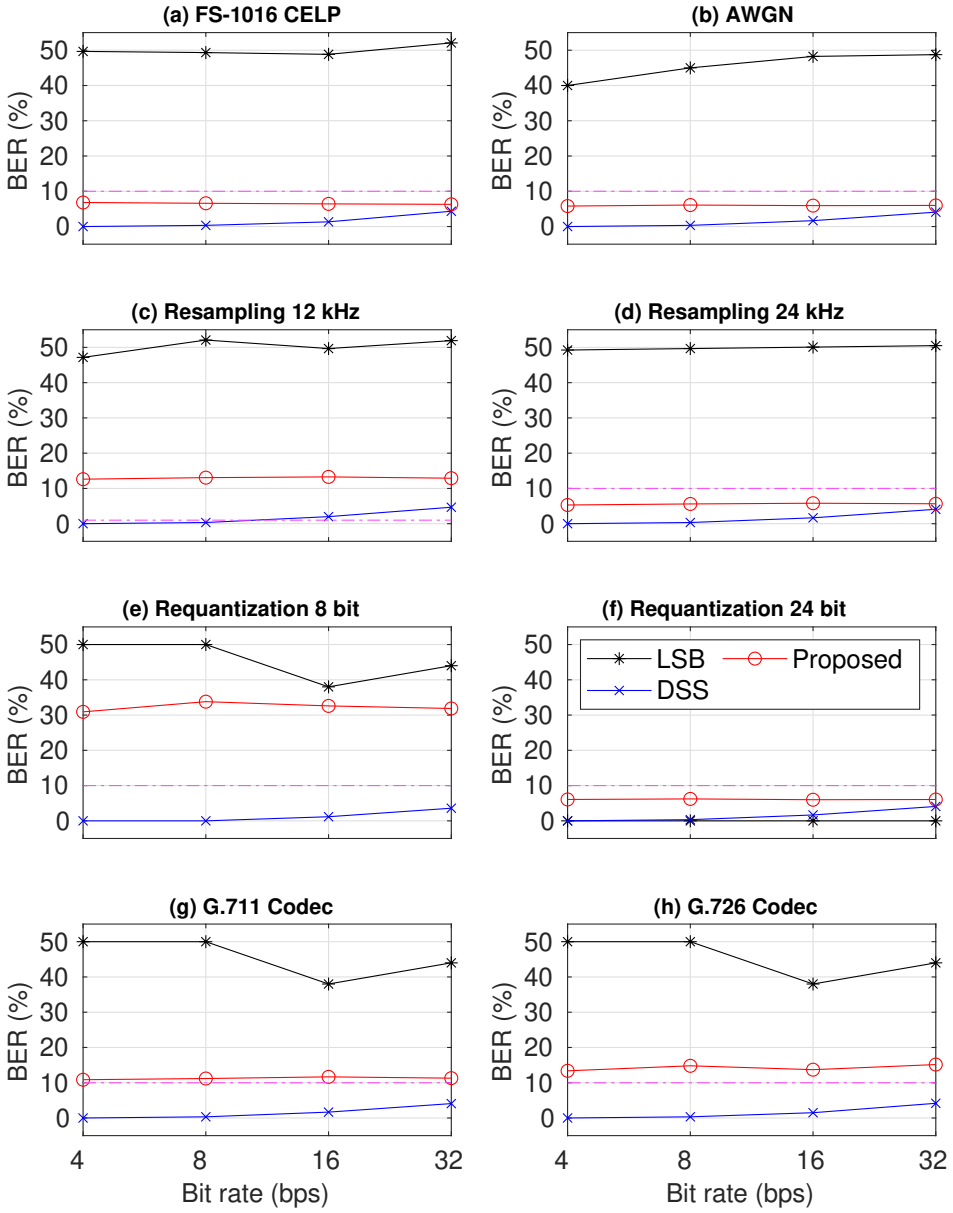

Fig. 9. Comparative robustness evaluation of our proposed method (single embedding), LSB, and DSS against signal processing attacks: (a) FS-1016 CELP codec, (b) Gaussian noise addition (AWGN), (c) down-sampling to 12 kHz, (d) up-sampling to 24 kHz, (e) requantization to 8 bit, (f) requantization to 24 bit, (g) G.711 codec, and (h) G.726 codec.

G.711 and G.726 (Fig. 9(g–h)).

V. Conclusion

This paper proposed a speech information hiding method on LSFs by modifying the LSB of the FS-1016 CELP quantization index. Our evaluation confirmed the successful embedding feasibility with regard to inaudibility and robustness under normal, resampling, noise addition, and speech codec (G.711, G.726, FS-1016 CELP) conditions. The evaluation results of single embedding showed that the least significant quantization bits of LSFs 1, 2, and 10 are more fragile; therefore, not good as embedding mediums. We also investigated two prospective ways to improve the robustness and payload by multiple embedding and varying the frame segmentation length. The results showed that multiple embedding could provide better robustness and higher payload, whereas reducing the frame segmentation could improve the payload with high accuracy at high bitrates but reduce the inaudibility.

As our future direction, we will improve the performance of the proposed method especially in dealing with attacks by

several signal processing operations. Moreover, we will consider other more advanced and frequently used CELP codecs, such as G.729 and AMR. The evaluation on the robustness of our proposed method against other frequently used speech communication attacks, such as frame synchronization and tampering, will also be conducted.

ACKNOWLEDGMENT

This work was supported by a Grant-in-Aid for Scientific Research (B) (No. 17H01761) and Grant-in-Aid for JSPS Research Fellow (No. 20J20580).

REFERENCES

[1] Z. Wu. 2014. Information Hiding in Speech Signals for Secure Communication (1st. ed.). Syngress Publishing.

[2] G. Hua, J. Huang, Y.Q. Shi, J. Goh, and V.L.L. Thing. 2016. Twenty years of digital audio watermarking-a comprehensive review. Signal Process. 128, C (November 2016), 222—242. DOI:https://doi.org/10.1016/j.sigpro.2016.04.005

[3] M. Unoki and D. Hamada. 2010. Method of digital-audio watermarking based on cochlear delay characteristics. J. Inn. Com. Inf., and Cont., 6, 3(B) (2010), 1325—1346.

[4] F. Djebbar, B. Ayad, Karim abed-meraim, and H. Hamam. 2012. Comparative Study of Digital Audio Steganography Techniques. EURASIP Journal on Audio, Speech, and Music Processing. (2012). DOI:https://doi.org/10.1186/1687-4722-2012-25

[5] S. Wang, W. Yuan, J. Wang, and M. Unoki. 2018. Speech Watermarking Based on Robust Principal Component Analysis and Formant Manipulations. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'18), (2018), 2082–2086. DOI:https://doi.org/10.1109/ICASSP.2018.8462356

[6] M.R. Schroeder and B.S. Atal. 1985. Code-excited linear prediction (CELP): High-quality speech at very low bit rates. IEEE International Conference on Acoustics, Speech, and Signal Processing (1985), 937-940. DOI:https://doi.org/10.1109/ICASSP.1985.1168147

[7] B.S. Atal and J. Remde. 1982. A New Model for LPC Excitation for Producing Natural Sounding Speech at Low Bit Rates. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'82), (1982), 614–617.

[8] Z. Wu. 2006. Speech Information Hiding in G.729. Chinese Journal of Electronics (ICJE) 15(3), (July 2006), 545–549.

[9] S. Wang and M. Unoki. 2014. Watermarking of speech signals based on formant enhancement. European Signal Processing Conference (EUSIPCO'14), (2014), 1257—1261.

[10] S. Wang and M. Unoki. 2013. Watermarking Method for Speech Signals Based on Modifications to LSFs. In Proceedings of the 2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP '13). IEEE Computer Society, USA, 283-–286. DOI:https://doi.org/10.1109/IIH-MSP.2013.79

[11] Z. Wu, W. Yang, and Y. Yang. 2003. ABS-based Speech Information Hiding Approach. IEEE Electronics Letters 39 (22) (2003), 1617-–1619.

[12] J.P. Campbell Jr., T.E.Tremain, and V.C. Welch. 1991. The Federal Standard 1016 4800 bps CELP Voice Coder. Digital Signal Processing, Academic Press, 1, 3 (1991), 145—155.

[13] K. Ramamurthy and A. Spanias. 2010. MATLAB Software for the Code Excited Linear Prediction Algorithm: The Federal Standard-1016. Morgan and Claypool Publishers.

[14] R. Salami et al. 1998. Design and Description of CS-ACELP: A Toll Quality 8 kb/s Speech Coder. IEEE Trans. on Speech and Audio Proc., 6, 2 (1998), 116–130. DOI:https://doi.org/10.1109/89.661471

[15] N. Sugamura and F. Itakura. 1981. Speech Data Compression by LSP Analysis/Synthesis Technique. Trans. IEICE, J64 (1981), 599-–606.

[16] I.V. McLoughlin. 2008. Review: Line spectral pairs. Signal Process. 88, 3 (March 2008), 448-–467. DOI:https://doi.org/10.1016/j.sigpro.2007.09.003

[17] F.K. Soong and B-H Juang. 1984. Line Spectrum Pair (LSP) and Speech Data Compression. IEEE International Conference on Acoustics, Speech, and Signal (ICASSP'84), (1984), 1.10.1–1.10.4.

[18] P. Kabal. 2003. Ill-Conditioning and Bandwidth Expansion in Linear Prediction of Speech. IEEE International Conference on Acoustics, Speech, and Signal (ICASSP'03), 1 (2003), I-824–I-827.

[19] A.J. Gray and J. Markel. 1976. Distance measures for speech processing. IEEE International Conference on Acoustics, Speech, and Signal (ICASSP'76), 24, 5 (1976), 380–391.

[20] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In Proceedings of the Acoustics, Speech, and Signal Processing, 200. on IEEE International Conference - Volume 02 (ICASSP '01). IEEE Computer Society, USA, 749-–752. DOI:https://doi.org/10.1109/ICASSP.2001.941023

[21] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano. 1990. ATR Japanese speech database as a tool of speech recognition and synthesis. Speech Communication 9 (1990), 357–363.