SPEECH BANDWIDTH EXTENSION BY DATA HIDING AND PHONETIC CLASSIFICATION

Siyue Chen and Henry Leung

University of Calgary
Department of Electrical and Computer Engineering
2500 University Drive NW, Calgary, Alberta, Canada, T2N 1N4

ABSTRACT

Speech bandwidth extension can be defined as the deliberate process of expanding the frequency range (bandwidth) for speech transmission. Its significant advancement in recent years has led to the technology being adopted commercially in several areas including psychoacoustic bass enhancement of small loudspeakers and the high frequency enhancement of perceptually coded audio. In this paper, a data hiding method based on dither quantization is used for speech bandwidth extension. More specifically, the out-of-band information is encoded and embedded into the narrowband speech without degrading the quality of the bandlimited signal. At the receiver, when the out-of-band information is extracted from the hidden channel, it can be used to combine with the bandlimited signal, providing a signal with a wider bandwidth. To encode the out-of-band speech more efficiently, acoustic phonetic classification is employed to generate three linear prediction (LP) codebook. The simulation results show that compared with using non-classified codebook, the propose scheme have a better bandwidth extension performance in terms of log spectral distortion (LSD).

Index Terms— Bandwidth extension, data hiding, acoustic phonetic classification, codebook mapping

1. INTRODUCTION

speech transmitted in communication networks is mostly in narrowband (NB) by using audio bandwidth (0.3-3.4 kHz) and the sampling frequency (8 kHz) originating from conventional pulse coding modulation (PCM). Both pleasantness and intelligibility of NB speech suffer from the limited bandwidth. Many efforts have been devoted to artificially generate WB speech from the NB speech. Conventional extension approaches [1, 2, 3] exploit the mutual dependence between NB and highband (HB) to estimate the missing HB signal from the NB signal. The estimated HB components are then used along with the NB speech to reconstruct a wideband (WB) speech. An accurate estimation of HB components usually requires a complicated speaker-dependent training of statistical models, which is computation-costive and thus not feasible for real-time processing. Although the training process can also be carried out off-line, i.e., speaker-independently, the performance of WB speech reconstruction degrades significantly.

Recently, data hiding [4, 5] has been proposed for speech bandwidth extension. More specifically, the encoded HB information is imperceptibly embedded into the NB speech by modulating the imperceptual components below the perceptual masking thresholds. While the hidden information can be retrieved from the received NB speech, it is used to reconstruct a WB speech with better quality and intelligibility. Compared to the conventional estimation-based ABE methods, the proposed data hiding scheme has an advantage of using the real HB information instead of the estimated one, thus is more accurate in reconstructing WB speech.

There are two challenges for the data hiding-based bandwidth extension schemes. On one hand, we want as much HB information as possible to be transmitted through the hidden channel. On the other hand, the hidden signal should be robust to noise corruption or quantization process. In our paper [4], we proposed to use linear prediction coding (LPC) and vector quantization (VQ) to encode a frame of HB signal into 23 binary bits. These 23 bits are then imperceptibly embedded into 80 NB samples by the data hiding method based on dither quantization. Although the proposed scheme outperforms conventional extension approaches in terms of the performance of WB speech reconstruction, it is also found that the proposed scheme degrads the perceptual quality of NB speech due to the large amount of distortion introduced by those 23 bits.

In this paper, we propose to use acoustic phonetic classification to further reduce the number of bits to be embedded. The simulation results demonstrate that only 12 bits are needed to achieve the same WB reconstruction performance as using 23 bits in [4]. Meanwhile, the perceptual quality of NB speech is improved significantly compared to that in [4]. The paper will start with the description of the data hiding scheme combined with acoustic phonetic classification. The hardware implementation of the proposed scheme is presented in Section 3. Experimental results are reported in Section 4. Finally, the conclusion remarks are presented.

2. DATA HIDING SCHEME WITH ACOUSTIC PHONETIC CLASSIFICATION

The flowchart of the proposed data hiding scheme is plotted in Figure 1. As shown, WB speech with the sampling rate of 16 kHz first undergoes band split by a low-pass and a high-pass filter respectively. The output of the low-pass filter is then down-sampled to provide the NB speech $x_{NB}(k)$, $0 \leq k \leq N-1$, where N is the number of speech samples. The output of the high-pass filter is shifted to the NB frequency range, and also decimated to provide an NB version of the HB speech, i.e., $x_{HB}(k)$, $0 \leq k \leq N-1$.

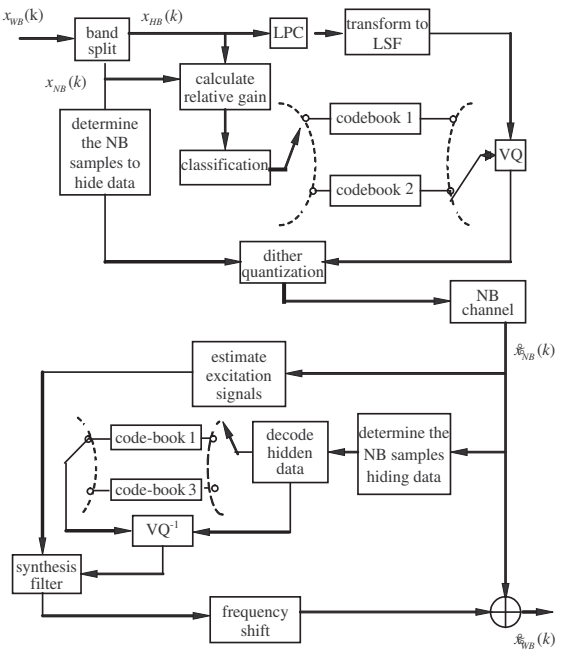

Fig. 1. The flowchart of the data hiding scheme.

In order to imperceptibly embed HB information into NB speech, it would be desirable to minimize the number of the digital bits that represent $x_{HB}(k)$. In this study, LPC [6], VQ and acoustic phonetic classification are employed to achieve this purpose. LPC is based on the assumption that human voice production can be modelled by a process of passing excitation signals through an all-pole synthesis filter. The filter coefficients are the reciprocal of the coefficients of an auto-regressive (AR) filter. Assume that the AR coefficients of $x_{HB}(k)$ are denoted as $\{a_{HB}(i), i = 1, \ldots, N_a$, where N_a is the filter order. They are obtained by using the Levinson-Durbin algorithm [6] to solve the set of equations

$$\sum_{i=1}^{N_a} a_{HB}(i)u'(|i-j|) = -u'(j), j = 1, \ldots, N_a, \quad (1)$$

where $u'(i)$ is the modified autocorrelation coefficients. Before applying VQ on the AR coefficients, they should be transformed to line spectral frequencies (LSF). This is because AR coefficients are sensitive to quantization errors. A slight change in AR coefficients will result in significant distortions when reconstructing HB speech. Instead, LSF are relatively less sensitive to quantization errors. Moreover, the AR coefficients can be precisely transformed back from the corresponding LSF.

Besides LSF, the gain of $x_{HB}(k)$ should also be embedded since the synthesized HB speech has to be scaled to an appropriate energy to avoid over-estimation [7]. Therefore, the relative gain of $x_{HB}(k)$ against $x_{NB}(k)$, i.e., $G_{rel} = \frac{G_{HB}}{G_{NB}}$, is calculated, and combined with N_a LSF to provide a representation vector of $x_{HB}(k)$, i.e., $\mathbf{a} = [LSF_1, \ldots, LSF_{N_a}, G_{rel}]$. $\mathbf{a}$ is then quantized to the closes entry of a VQ codebook that is generated by the fuzzy c-means (FCM) algorithm [8]. By doing so, only the entry index, instead of $\mathbf{a}$, is to be embedded into NB speech. It is noted that two codebooks are used in VQ. One is for the situation of $G_{rel} < -15$ dB, and the other is for $G_{rel} \geq -15$ dB. This is because that the situation of $G_{rel} \geq -15$ dB has a low probability to occur. If only one codebook is used, the representation vectors belonging to this category may all quantized to one entry, which will result in significant information loss.

Assuming that the codebook size is N_c, $\log_2 N_c$ binary digits are needed to represent a specific entry index. Considering one more bit is required to indicate which codebook is used, we then have to embed $\log_2 N_C + 1$ data bits in total, i.e., $\{b_m\}, m = 0, 1, \ldots, \log_2 N_c + 1$. The process of embedding b_m into the imperceptible components are formulated as

$$x'_{NB}(k) = Q(x_{NB}(k) + q_m) - q_m, \quad (2)$$

where q_m takes a value following the rule that

$$q_m = \begin{cases} \frac{\Delta_m}{4} & \text{if} \quad b_m = 1 \\ -\frac{\Delta_m}{4} & \text{if} \quad b_m = 0 \end{cases}, \quad (3)$$

and $Q(\cdot)$ denotes the quantization process with the step size Δ_m. The $\log_2 N_C + 1$ speech samples with the least amplitude are chosen to hide the encoded bits $\{b_m\}$, since the modification on them is less imperceptible. The value of Δ_m is determined by the step size used in A-law or μ-law to quantize $x_{NB}(k)$. For example, if μ-law is used, and the speech sample that b_m is embedded has $\mu \frac{x_{NB}(k)}{x_{max}} \ll 1$, Δ_m can be set as

$$\Delta_m = (1+\alpha)\Delta_\mu \frac{x_{max} \ln \mu}{y_{max}\mu}, \quad (4)$$

where x_{max} and y_{max} indicates the input and the output ranges of the μ-law compander, $\mu = 255$ when 8 bits are used to represent one speech sample, Δ_μ is the step size with which the output of the μ-law compander is uniformly quantized. Interested readers can go to [9] for more details about these parameters. Only α is not related to μ-law, but a parameter within the range $(0, 1)$ to achieve a tradeoff between robustness and imperceptibility of the hidden data. A large value

of α is desirable for the robustness consideration, however, also results in more embedding distortion. Therefore, we can choose a value slightly over 0.

The NB signal, x'_{NB}, is transmitted through the NB channel. It might be corrupted by processing or channel noises. Let's denote the received signal as $\tilde{x'}_{NB}$. A conventional NB equipment will treat $\tilde{x'}_{NB}$ as an ordinary speech signal. Since the difference between x'_{NB} and x_{NB} is so small that it is almost imperceptible, the perceptual quality of the NB speech will not be significantly degraded. Meanwhile, if frame synchronization and partition can be achieved, the proposed scheme can extract the hidden data by applying a minimum-distance decoder, i.e.,

$$\tilde{q}_m = \min_{\tilde{q_m} \in \left\{ \frac{\Delta_m}{4}, -\frac{\Delta_m}{4} \right\}} \left| \tilde{x'}_{NB}(k) - Q(\tilde{x'}_{NB} + \tilde{q}_m) + \tilde{q}_m \right|, \tag{5}$$

and then making a decision by

$$\tilde{b}_m = \begin{cases} 1 & \text{if} \quad \tilde{q_m} = \frac{\Delta_m}{4} \\ 0 & \text{if} \quad \tilde{q_m} = -\frac{\Delta_m}{4} \end{cases} . \tag{6}$$

Given b_m, the quantized LSF and the relative gain can be properly retrieved from the VQ codebook. The LSF are transformed back to the AR coefficients. Meanwhile, the excitation signal is obtained as the residual of an LPC analysis on the received NB signal, i.e., $r(k) = \hat{x}_{NB}(k) - \sum_{i=1}^{N_a} a_{NB}(i)\, \hat{x}_{NB}(k-i)$, where $r(k)$ is the residual and $a_{NB}(i)$ denotes the AR coefficients of the received NB speech. $r(k)$ is then used to excite the all-pole synthesis filter described by the coefficients obtained from the quantized LSF, generating $\hat{x}_{HB}(k)$, the reconstructed $x_{HB}(k)$. The gain of $\hat{x}_{HB}(k)$ is adjusted to $\hat{G}_{HB}$, which is obtained by $\hat{G}_{HB} = \hat{G}_{NB} \cdot \bar{G}_{rel}$, where $\hat{G}_{NB}$ is the gain of the received NB speech, $\bar{G}_{rel}$ is the one retrieved from the VQ codebook. At this point, $\hat{x}_{NB}(k)$ and $\hat{x}_{HB}(k)$ are still the signals sampled at 8 kHz. They should be up-sampled to 16 kHz, the sampling rate of WB speech. In addition, the up-sampled $\hat{x}_{HB}(k)$ should be shifted to its destination frequency band by a high-pass filter, providing the restored HB speech. Finally, a WB speech is artificially generated by adding the restored HB speech to the up-sampled NB speech.

3. HARDWARE IMPLEMENTATION

Hardware implementation of the proposed data hiding scheme is presented in this section. The processor used in our design is a Xilinx MicroBlaze soft processor. It is a reduced instruction set computer (RISC) CPU, which allows users to select any combination of peripherals and controllers. Furthermore, it can be easily connected with user-defined hardware acceleration intelligent property (IP) core through its on-chip peripheral bus (OPB) or its peripheral local bus (PLB). This feature helps to maximally meet the requirements of a non-standard design. The FPGA employed in this study is Virtex II XCV2000FF896-4, a ball grid arrays (BGA) packed 624-pin large memory resource chip. When synthesizing the processor into the FPGA, the clock rate is determined at 84.545 MHz.

Figure 2 plots the hardware design of the proposed data hiding scheme . As shown, a single soft processor is used for both encoding and decoding due to the limitation of hardware resources. The encoder output $x'_{NB}(k)$ is thus directly connected to the decoder input. Since the sampling rates of $x_{NB}(k)$ and $x_{HB}(k)$ are 8 kHz and the clock rate is 84.545 MHz, three FIFOs (First in, First out) are used in both the input and the output port so that the fast processing speed are able to match the low sampling rate. Each FIFO has its word width as 8 bits and its depth as 512. In total, three FIFOs consume 1.5 kilobytes of the FPGA resource. The FIFOs are connected with the processor through PLB, which can read an 8-bit data within one clock period.

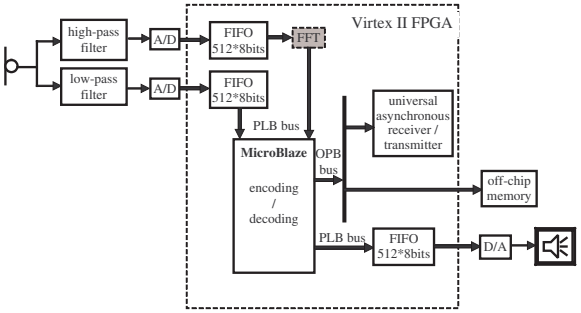

Fig. 2. Hardware implementation of the data hiding scheme.

Hardware performance is usually evaluated by resource consumptions. The less the resources are consumed, the better the hardware performance. Table 1 shows the hardware performances of the proposed data hiding scheme. The used LUTs are 4356 out of 21504, leading to the usage rate as 20.3%. In addition, the usage rates for slice and flip-flop are 17% and 8% respectively. Hence, only a small portion of resources are needed to implement the data hiding scheme. In the standard speech communication system, the one-way speech transmitting delay can not exceed 200 ms per frame of 8000 speech samples in order to achieve comfortable voice communication. For this scheme, 0.422 ms encoding-decoding time is able to meet this requirement.

4. EXPERIMENTAL RESULTS

The assessment of the perceptual quality for the composite signal x'_{NB} is carried out by mean opinion score (MOS). The subjects are asked to compare x_{NB} with x'_{NB} and give their opinions. α is set as small as 0.05 to reduce embedding distortion. Meanwhile, we have $\log_2 N_C + 1 = 12$. Scaling of MOS is 4 grades and their instructions are as follows:

Table 1. Hardware performances of data hiding scheme

processing time (ms/per b_i)	encoding: 0.212 decoding: 0.21
logic consumption	slice: 1621 flip-flop: 1405 LUT: 2165
memory consumption (kBytes)	148588
power consumption (mW)	436

1: Two signals are quite different

2: Two signals are similar, but the difference is easy to see

3: Two signals are very similar, only little difference exists

4: Two signals sound identical

The two signals are played in a random order, and the sound pressure level is set as 63 dB. Obtained by averaging ratings of all subjects over all testing signals, the resultant score comes to 3.87. Therefore, the perceptual quality of the original NB speech is well maintained.

To evaluate the perceptual similarity of the reconstructed and the original HB speech signal, LSD defined as

$$LSD = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(20\log_{10}\frac{G}{|A_{HB}(e^{j\omega})|} - 20\log_{10}\frac{\tilde{G}}{|\tilde{A}_{eb}(e^{j\omega})|}\right)^2 d\omega, \quad (7)$$

where $A_{HB}(e^{j\omega}) = \sum_{i=1}^{10} a_{HB}(i)e^{-ji\omega}$ and $G = G_{HB}/G_{NB}$, is still used as the performance measure due to its reasonable correlation with the subjective speech quality. For the comparison purpose, we also compute the LSD results of using the data hiding method in [4]. It was found that LSD of the proposed scheme in this paper is 1.93 dB less than that of the scheme in [4]. Therefore, although the amount of data that represent HB speech is significantly reduced, the proposed scheme here has a better WB reconstruction performance.

5. CONCLUSION

This paper proposes a data hiding scheme to extend speech bandwidth. More specifically, HB speech is encoded into digital bits and embedded into NB speech imperceptibly. When the hidden data is decoded at the receiver, HB speech can be reconstructed and combined with NB speech to provide a WB speech. It is shown that the proposed scheme has a good WB reconstruction performance in terms of LSD. Furthermore, it has a processing speed of 0.422 ms for encoding-decoding a frame of 80 NB speech samples. Compared to the data hiding scheme in [4], the scheme in this paper is able to use much less data bits to encode HB speech, thus can achieve better performance of reconstructing WB speech in terms of LSD. Meanwhile, the perceptual quality of NB speech is well maintained. Compared to the HMM codebook mapping method, the proposed scheme is able to meet the requirement of real-time processing and consumes less hardware resources. It provides a practical and effective solution to speech bandwidth extension.

6. REFERENCES

[1] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, pp. 1707–1719, 2003.

[2] J. Kontio L. Laaksonen and P. Alku, "Artificial bandwidth expansion method to improve intelligibility and quality of amr-coded narrowband speech," in *Proc. IEEE ICASSP*, Philadelphia, March 2005, vol. 1, pp. 809–812.

[3] T. Unno and A. McCree, "A robust narrowband to wideband extension system featuring enhanced codebook mapping," in *Proc. IEEE ICASSP*, Philadelphia, USA, March 2005, vol. 1, pp. 805–808.

[4] S. Chen and H. Leung, "Artificial bandwidth extension of telephony speech by data hiding," in *Proc IEEE ISCAS*, Kobe, Japan, May 2005, pp. 3151–3154.

[5] H. Ding, "Sub-channel below the perceptual threshold in audio," in *Proc IEEE ICASSP*, April 2003, vol. 2, pp. 325–328.

[6] F. C. A. Somerville L. Hanzo and J. P. Woodard, *Voice Compression and Communications: Principles and Applications for Fixed and Wireless Channels*, IEEE Press, 2001.

[7] M. Nilsson and W. B. Kleijn, "Avoiding over-estimation in bandwidth extension of telephony speech," in *Proc. IEEE ICASSP*, May 2001, pp. 869–872.

[8] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, New York: Plenum, 1981.

[9] N. S. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*, Englewood cliffs, NJ: Prentice-Hall, 1984.