WIDEBAND AUDIO OVER NARROWBAND LOW-RESOLUTION MEDIA

Heping Ding, ***Senior Member, IEEE***
Institute for Microstructural Sciences, National Research Council, Canada
heping.ding@nrc-cnrc.gc.ca

ABSTRACT

This paper proposes a scheme of transmitting or storing a wideband audio stream in a channel or storage medium which is narrowband in nature and with a low resolution data format, such as an 8-bit companded one. In the proposed scheme, signal components in the upper-band - part of the wideband beyond the narrowband - are encoded and then the code is embedded into the narrowband stream in a least audible way. As a result, this scheme enables the transmission and storage of wideband audio signals with the existing narrowband infrastructure. Furthermore, the scheme is fully backward compatible, i.e., conventional receivers, without the proposed decoding mechanism, can still access the narrowband audio with little audible degradation. The proposed scheme can find applications in areas where a high quality wideband audio is needed but the physical bandwidth is limited. Examples are voice-over-IP and telephony in general, as well as digital storage and playback. An audio demonstration of the proposed scheme will be available at the presentation.

1. INTRODUCTION

A TELEPHONY voice channel, such as one with the standard public switched telephone network (PSTN), digital private branch exchange (PBX), or voice-over-IP (VoIP), is only able to transmit audio signal components up to about 3400 Hz. A signal so transmitted is regarded as of "narrow frequency band" (NB), as illustrated by the solid curve in Fig. 1.

Such a small bandwidth makes a telephony channel only able to transmit "toll quality" voices, with intelligibility and subjective quality much inferior to that of the so-called wideband (WB) voice, whose bandwidth is 50 - 7000 Hz as defined by International Telecommunication Union (ITU). As shown in Fig. 1, the WB is NB extended by an upper-band (UB), indicated by the dashed curve.

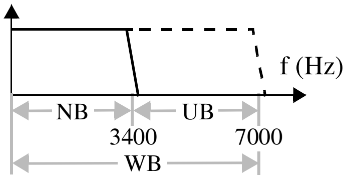

Fig. 1. Bandwidth definitions

With NB being a nature of the existing telephony infrastructure, we cannot count on an easy way of increasing the physical bandwidth without prohibitively increasing the cost.

Being a continuation of [1], this paper proposes a scheme that virtually, as opposed to physically, extends an NB channel's bandwidth so that the ITU wideband (WB) voices can be transmitted over the existing NB infrastructure. In addition, the proposed scheme remains compatible with the existing NB infrastructure.

Note that, in this paper, we only consider extending the bandwidth of an NB channel at the high-end, i.e., beyond 3400 Hz. This is because the transmission at the low frequency end is usually not a problem in a digital network.

The rest of the paper is organized as follows. Section 2 reviews the existing approaches in the field. The goal of our research is discussed in Section 3. The proposed scheme is discussed in Section 4, and some results are given in Section 5. Finally, Section 6 is a summary.

2. EXISTING APPROACHES

Certain efforts have been made to extend the virtual bandwidth of an NB channel. Existing approaches can be classified into certain categories below.

Voice and audio coding Wideband voice coding schemes encode wideband voice signals into digital bits. An audio coding scheme encodes audio signals in general into digital bits. The digital information can then be transmitted over an NB channel. Typical wideband voice coders are [2][3][4], and a typical audio coder is [5]. Such a configuration transmits a signal that cannot be understood to the human ear without a proper decoding scheme. In other words, it is incompatible with the conventional end-user audio equipment.

Simultaneous voice and data (SVD) Being capable of transmitting data in addition to voice over a PSTN channel so as to increase the capacity of the latter, SVD technology is often used in dial-up modems that connect computers to the Internet through the PSTN. Typical examples are in [6] and references therein. SVD approaches change the audio signal significantly and need SVD-capable modem hardware; therefore, they are not directly compatible with the conventional end-user audio equipment and are costly.

Audio watermarking These techniques increase the capacity of an NB channel by embedding information in an audio stream in ways so that it is inaudible to the human ear. An overview of the technology can be found in [7]. These techniques are aimed at

0-7803-8484-9/04/$20.00 ©2004 IEEE

high security, i.e., low probability of being detected or removed by a potential attacker, and low payload rate. Our requirements for extending the NB media's capacity are just the opposite; we want a high payload rate, while the security is considered less of an issue.

3. OBJECTIVES

Our objectives for a scheme that is able to transmit wideband voice over an NB channel are as follows.

Simplicity Firmware implementation should be simple and extra hardware should be none or minimal.

Compatibility with the existing end-user equipment A conventional NB phone terminal should still be able to access the basic NB voice service as usual, although it is not able to let the user enjoy the wideband audio. This feature

- is useful in audio broadcasting and conferencing applications, where the audience may consist of a mixture of ordinary NB terminals and ones equipped with the wideband decoding mechanism, and
- will greatly facilitate the phase-in of this new technology.

High payload rate It should be higher than that offered by audio watermarking schemes while the stringent security requirement incurred by them can be eased. That is, the additional payload could possibly be tempered by an attacker. However, an attacker is not necessarily able to obtain the information therein if an encryption scheme is used.

4. PROPOSED SCHEME

The proposed scheme encodes signal components in the upper-band (UB), as shown in Fig. 1, and embeds the encoded information into the NB signal. In doing so, there are two challenges:

- the UB signal should be encoded with as few bits as possible, in order to minimize the information that has to be embedded into the NB signal; and
- when in applications with 8-bit companded data formats, i.e., μ-law and A-law specified in [8], the scheme must not create any annoying artifacts. This is because, compared with a higher-resolution data format such as the 16-bit linear one, an 8-bit companded data format already has an audible noise floor and further manipulations to the data will easily worsen the problem.

The scheme consists of a transmitter and a receiver. They are discussed below.

4.1. Transmitter

The proposed transmitter, shown in Fig. 2, partitions the original WB audio stream, with a sampling rate of 16 kHz, into non-overlapped *N*-sample frames and processes them one after another. In our exercise, *N*=160 is chosen so that the frame size is 10 ms. It takes the following steps to process each frame.

Band split The samples in the frame are filtered by two filters, being low-pass and high-pass which produce two outputs, *NB* and *UB*, respectively.

***UB* down-shift** The *UB* output of the last step undergoes a frequency down-shift operation. As a result, a frequency shifted version of *UB*, UB_s fits into the NB frequency range, as shown in Fig. 2.

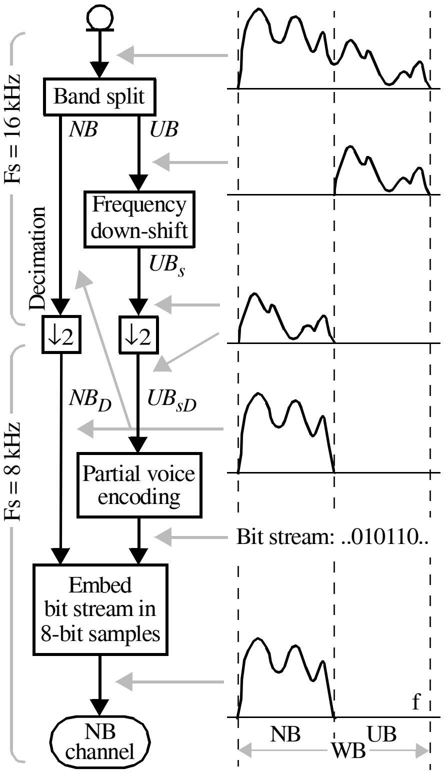

Fig. 2. Proposed transmitter

***NB* and UB_s decimation** Since both *NB* and UB_s are band-limited to below 4000 Hz, they can be down-sampled, i.e., decimated, so that the sampling rate for them is halved, to 8000 Hz., which is the NB channel's sampling rate.

UB_{sD} encoding UB_{sD}, the frequency-shifted and decimated version of the upper band signal *UB* is encoded into digital bits. This can be done with the use of a standard encoding scheme or a part of it. In our prototypes, we have tried two methods with pretty good results. One of the schemes we have tried is to encode the linear predictive coding (LPC) coefficients and gain. In particular, part of the ITU-T G.729 [9] NB voice codec is used. Every 10 ms, or 80 NB data samples, the parameters transmitted by a full G.729 encoder consists of those for LPC coefficients (18 bits) and those for the faithful regeneration of the excitation at the decoder (62 bits), totalling 80 bits per frame or 1 bit per data sample. Therefore, if we used a full G.729 scheme to code UB_{sD}, we would have to modify up to one bit from each NB data sample. This, as discussed earlier, would likely cause unacceptable noise with an 8-bit companded data format.

Alternatively, in our proposed approach, the excitation signal is not encoded at the transmitter; instead, it is derived at the receiver as the residue of an LPC analysis on the received NB signal. This method reduces the number of bits to be transmitted and the system complexity; it does not need any explicit voiced/unvoiced decision and control. This is because the LPC residue so derived will automatically be periodic-like when the received NB signal is voiced, and white-noise-like when the signal is unvoiced. As a result, our encoding/decoding scheme for the upper-band is much simpler than a vocoder, and we only need to transmit the parameters for the LPC coefficients (18 bits required with G.729) and no more than 5 bits for the gain - totaling 23 bits per 80-sample frame, as opposed to 80 required by the full G.729.

Embedding bit stream into NB samples the 23 bits representing the encoded UB_{sD} are to be embedded into 80 NB_D samples, with the format being 8-bit companded, as per [8] and shown in Fig. 3. The embedding is done in a way to minimize the added noise. First, the frame of 80 samples is partitioned into 23 groups. Being the host to embed 1 bit, each group contains 3 or 4 8-bit samples, or group members. Next, we need to modify the mantissa of one group member in order to embed the bit. One (not necessarily the only) way of doing this is to pick the group member with the smallest magnitude, and to modify its M_0 as per Table 1.

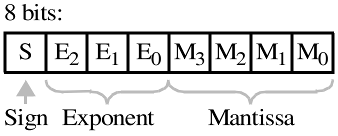

Fig. 3. μ-law or A-law format

Bit to be embedded	Algebraic sum of group members	M_0 of group member with smallest magnitude
0	Even	No modification
	Odd	Flip (0 ↔ 1)
1	Even	Flip (0 ↔ 1)
	Odd	No modification

Table 1. M_0 modification scheme

As a result of this operation, the sum of the group members will be even if the embedded bit is 0, and odd otherwise. It can be proven that the modification is unbiased; it does not distort the signal but adds noise.

Synchronization During frames where there is no audio activity, a unique 23-bit pattern can be sent. These frames will help the receiver acquire and keep in synchronization with the transmitter.

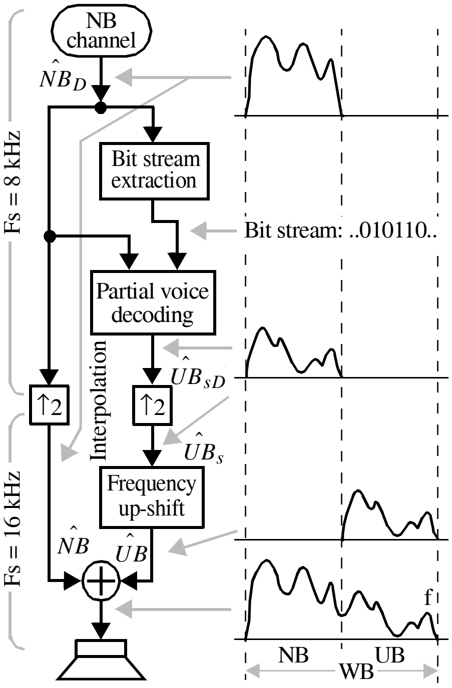

Fig. 4. Proposed receiver

4.2. Receivers

A conventional digital receiver, being NB, treats the received signal as an ordinary digital audio signal and sends it to an electro-acoustic transducer. The modifications made to certain M_0's have a minor impact on the perceptual audio quality and therefore will not significantly worsen the noise issue with an 8-bit companded format.

Once frame synchronization and partitioning have been properly looked after, the proposed receiver, shown in Fig. 4, takes the following steps to process each frame.

Bit stream extraction First, an 80-sample frame is partitioned into 23 groups the same way as that in the transmitter. Next, the sum of the 8-bit samples in each group is found. Last, the value of the bit embedded in each group is determined based on the parity of the sum.

Voice decoding In this step, we need to derive an excitation from the received NB signal and use it to excite an all-pole speech production model whose parameters are obtained by decoding the bits received. The excitation is actually the residue of an LPC analysis on the received NB signal. For fast convergence in order to obtain a well whitened residue, a fast converging adaptive lattice LPC filter [10] is used.

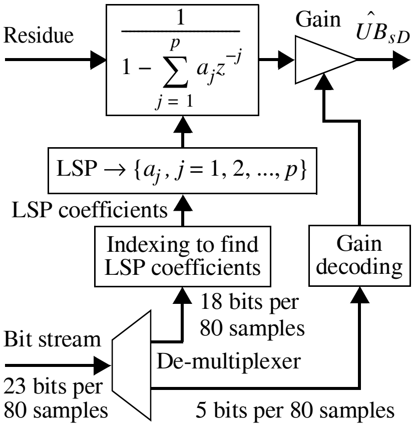

Fig. 5. Decoding for $\hat{UB}_{sD}$

The LPC residue is then used to excite an all-pole speech production model and the gain is properly adjusted, as in Fig. 5. We use part of the ITU-T G.729 decoder to decode the all-pole model coefficients $\{a_j, j = 1, 2, ..., p\}$ and to implement the all-pole implementation. However, this is not necessarily the case; another vocoding scheme can be used instead.

Interpolation for $\hat{NB}$ and $\hat{UB}_s$ At this point, $\hat{NB}_D$ and $\hat{UB}_{sD}$ are sampled at 8 kHz and they should be up-sampled to 16 kHz, the sampling rate of the final output.

$\hat{UB}_s$ up-shift This step is to move the decoded upper-band signal $\hat{UB}_s$, now occupying the NB, to its destination frequency band, i.e., the upper-band. The amount of frequency up-shift is the same as that of the frequency down-shift performed in the transmitter.

Summation to form output In this last step, the up-sampled received signal $\hat{NB}$ and the restored upper-band signal $\hat{UB}$ are added to form the WB output.

4.3. Complexity

From the above discussions, It can be seen that it is simpler to implement the proposed codec scheme than it is with the G.729, because the former only uses part of the latter and the rest of the algorithm is quite simple. In addition, there is a potential to use an order-reduced version of the partial G.729, which was meant to code the NB signal, to code the UB, so as to further reduce the complexity and the number of bits transmitted.

5. SOME COMPARISON RESULTS

So far, only preliminary and limited subjective evaluations have been conducted to compare the performance of the proposed scheme with that of G.722 [2] WB codec operating at different data rates.

Since the presence of the so-called idle channel noise associated with a G.722 codec would make a listener easily distinguish it from the proposed scheme, which has a lower noise floor, the G.722 output while there is no voice activity has been attenuated by 6 dB.

A total of 4 subjects listened to 4 Harvard sentences, spoken by 2 male and 2 female speakers. The sentences were processed by G.722 operating at 64, 56, and 48 kbits/s, as well as by our proposed scheme. The subjects were asked to rate the 4 outputs of each sentence as 1 (best), 2, 3, and 4 (worst).

Obtained by averaging over ratings of all subjects, the resultant scores are shown in Fig. 6. It can be seen that the proposed scheme is superior to G.722 at 48 kbits/s, and comparable in subjective quality to the 56 kbits/s version of G.722.

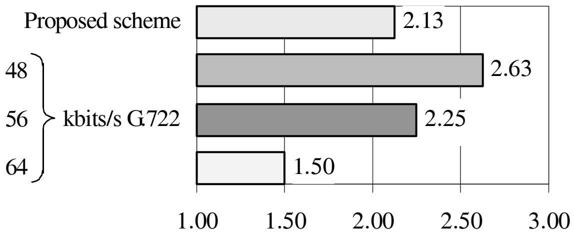

Fig. 6. Subjective evaluation scores

6. SUMMARY

In this paper, we have proposed a scheme that can transmit a WB voice over NB channels, or store a WB voice in NB storage media, where the channels or the media may have a low-resolution data format, such as the 8-bit companded one. Furthermore, the scheme is backward compatible; it transmits a signal that can be directly played as a conventional NB signal on an ordinary sound playing device - without the need for any processing.

The comparison between the proposed scheme and the ITU-T G.722 codec is summarized in Table 2.

In a word, as an alternative to a G.722 codec, the proposed scheme offers the same level of sound quality at the same data rate, and is compatible with the existing NB infrastructure. This feature can be particularly useful in simplifying audio broadcasting and conferencing applications and will greatly facilitate the phase-in of this new technology.

Codec scheme	**Data rate (kbits/s)**	**Subjective quality**	**Backward compatibility**
G.722	64	Very good	No
	56	Good	
	48	Acceptable	
Proposed	64	Good	Yes

Table 2. Comparison between G.722 and the proposed scheme

REFERENCES

[1] Heping Ding, "Sub-Channel Below the Perceptual Threshold in Audio," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, April 2003.

[2] International Telecommunication Union, "7 kHz AUDIO - CODING WITH 64 KBITS/S," ITU-T Recommendation G.722, 1993.

[3] International Telecommunication Union, "Coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss," ITU-T Recommendation G.722.1, 1999.

[4] International Telecommunication Union, "Wideband coding of speech at around 16 kbit/s using Adaptive Multi-rate Wideband (AMR-WB)," ITU-T Recommendation G.722.2, Jan. 2002.

[5] ISO/IEC JTC 1/SC 29/WG 11, ISO/IEC 13818-3 "Information technology - Generic coding of moving pictures and associated audio information - Part 3: Audio," (MPEG-2) April 15, 1998.

[6] Gordon Bremer, *et_al*, "Simultaneous Analog and Digital Communications with a Selection of Different Signal Point Constellations Based on Signal Energy," U. S. Patent 5,436,930, July 25, 1995.

[7] Ingemar J. Cox, *et_al*, *Digital Watermarking*, ISBN 1-55860-714-5, Academic Press, 2002.

[8] International Telecommunication Union, "PULSE CODE MODULATION (PCM) OF VOICE FREQUENCIES," ITU-T Recommendation G.711, 1972.

[9] International Telecommunication Union, "CODING OF SPEECH AT 8 kbits/s USING CONJUGATE-STRUCTURE ALGEBRAIC-CODE-EXCITED LINEAR-PREDICTION (CS-ACELP)," ITU-T Recommendation G.729, March 1996.

[10] Heping Ding, Chongzhi Yu, "Adaptive Lattice Noise Canceller and Optimal Step Size," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 2939-2942, Tokyo, Japan, April 1986.