A SCALABLE BANDWIDTH EXTENSION ALGORITHM

Visar Berisha and Andreas Spanias

Arizona State University
SenSIP Center, Department of Electrical Engineering
Tempe, AZ 85287
Email: [visar, spanias]@asu.edu

ABSTRACT

Most modern bandwidth extension techniques predict the high- frequency band based on features extracted from the lower band. While this works for some frames, problems arise when the correlation between the low and the high band is insufficient. In these situations, additional high-band information must be sent to the decoder. In this paper, we propose a scalable speech coding method based on the principles of bandwidth extension. The rate selection is based on explicit psychoacoustic criteria, while the bandwidth extension is performed using a constrained MMSE estimation technique. Objective and subjective evaluations indicate that the proposed system performs at a lower average bit rate when compared to other similar algorithms while improving speech quality.

Index Terms— bandwidth extension, multirate coding, scalable speech coding, speech enhancement, psychoacoustics

1. INTRODUCTION

The public switched telephony network and most of today's cellular networks use speech coders operating with a limited bandwidth (0.3 - 3.4 kHz). This in turn places a limit on the naturalness and intelligibility of the synthesized speech [1]. This is most problematic for sounds whose energy is spread over the entire spectrum. For example, unvoiced sounds such as 's' and 'f' are often difficult to differentiate with a narrowband representation. To combat the problem, algorithms that aim to recover a wideband (0.3 - 7 kHz) speech signal from its narrowband (0.3 - 3.4 kHz) counterpart have recently gained popularity [2] - [6]. Researchers have used estimation methods based on the underlying speech production model to restore the missing bands. In [2], vector quantization of model parameters is used at the decoder to reconstruct the wideband signal. In [3] a Gaussian mixture model and a hidden Markov model (HMM) are used to predict model parameters. The assumption for these approaches is that there is a sufficient correlation between the narrowband features and the wideband envelope. While this is true for some frames, the assumption does not hold generally [7]. In Fig. 1, we show examples of two frames that illustrate this. The figure shows two frames of wideband speech along with the true envelopes and predicted envelopes. The estimated envelope was predicted using a technique solely based on pre-trained Gaussian mixture models [5] [20]. The top figure shows a frame for which the predicted envelope matches the actual envelope quite well. In the bottom figure, the estimated envelope greatly deviates from the actual and in fact introduces two high-band formants. In addition, it misses the two formants located between 4 kHz and 5.5 kHz. To avoid this, one can encode the high band separately using only a limited number of parameters. Since the higher frequency band is perceptually less relevant than its low frequency counterpart, a coarse representation is often sufficient for a perceptually lossless representation [13]. This idea is also used in high-fidelity audio coding based on spectral band replication [15], where an underlying perceptual codec encodes the lower frequency band while the high band is coarsely parameterized using fewer parameters.

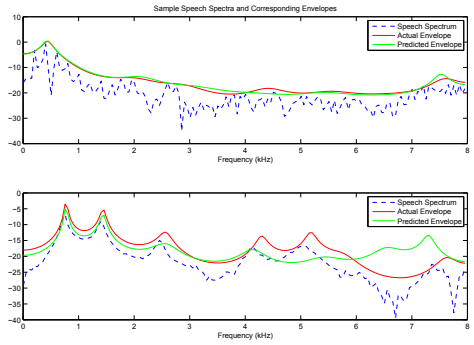

Fig. 1. Wideband speech spectra (in dB) and their actual and predicted envelopes

In this paper, we propose a multirate bandwidth extension algorithm that allocates bits only to frames that benefit from a wideband representation. A coder/decoder structure is proposed in which the lower frequency band is encoded using an existing linear prediction coder while the high band is generated using a novel method based on a constrained MMSE estimator. A rate determination algorithm based on explicit psychoacoustic criteria determines the appropriate mode for the coder on a frame-by-frame basis. The bandwidth extension algorithm is based on a source-filter model in which the high-band envelope and excitation are estimated separately. Depending upon the selected rate, the envelope and excitation are either parameterized at the encoder or predicted at the decoder. We compare the proposed scheme to the adaptive multi-rate (AMR) coder and show that the proposed algorithm achieves improved audio quality at a lower average bit rate.

This paper is organized as follows: Section 2 provides a description of the coder/decoder. In Section 3, sample results are presented and section 4 contains concluding remarks.

This work is supported by an NSF Graduate Research Fellowship. A patent pre-disclosure has been filed.

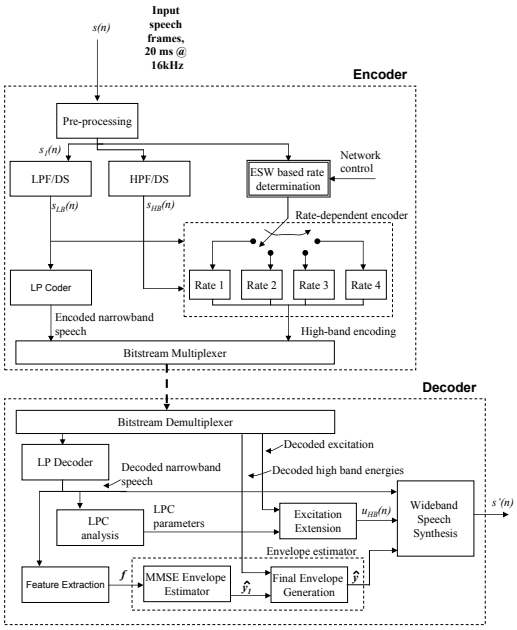

Fig. 2. A high level overview of the proposed encoder/decoder

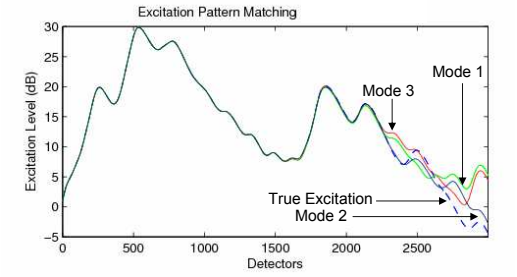

Fig. 3. The original excitation pattern and the excitation pattern associated with a frame encoded in 3 of the modes. For this frame, no extra information was deemed necessary

2. PROPOSED METHOD

A high level diagram of the proposed system is shown in Fig. 2. The algorithm operates on $20ms$ frames sampled at 16 kHz denoted by $s(n)$. The low band of the audio, $s_{LB}(n)$, is encoded using an existing linear prediction (LP) coder, while the high band, $s_{HB}(n)$, is artificially extended using an algorithm based on the source/filter model. The rate determination algorithm selects one of four possible modes under which the bandwidth extension system can operate. Depending upon the chosen rate at the encoder, the high-band excitation and envelope are appropriately parameterized and transmitted to the decoder. The decoder uses a series of prediction algorithms to generate estimates of the high-band envelope and excitation, respectively denoted by $\hat{\mathbf{y}}$ and $u_{HB}(n)$. These are then combined with the LP coded lower band to form the wideband speech signal, $s^{'}(n)$.

In this section, we provide a detailed description of the three main components of the algorithm: i) the rate determination algorithm ii) the envelope extension algorithm iii) the excitation extension algorithm.

2.1. Rate Determination

The inclusion of explicit psychoacoustic criteria in speech coding has been shown to remove perceptual redundancies [8] [9]. In this section we describe the rate determination algorithm based on excitation pattern matching and partial loudness [13] [12]. The proposed bandwidth extension algorithm operates in four possible modes outlined in Table 1. The purpose of the rate determination algorithm is to determine whether there is a perceptual gain in encoding the excitation and/or the envelope on a frame-by-frame basis. The excitation is encoded using the algebraic codebook found in the adaptive multi-rate (AMR) codec [11], whereas the envelope is encoded by transmitting energy values for subbands of the high-band.

The appropriate mode for the frame is selected using a determination - by - synthesis algorithm. At the encoder, the frame is coded/decoded using each of the four rates and the perceptual difference between the decoded frame and the true frame is analyzed. This difference is based on the difference in excitation patterns between the predicted speech and actual speech [14]. This difference is then compared to a threshold estimated using the metrics obtained in the previous 5 frames. To determine the number of subbands to encode, L, the authors refer to a previously proposed algorithm based on partial loudness [10].

	Modes of Operation	Required Bits
1	Excitation Predicted, Envelope Predicted	0 + 0
2	Excitation Encoded, Envelope Predicted	72 + 0
3	Excitation Predicted, Envelope Encoded	0 + $3L$
4	Excitation Encoded, Envelope Encoded	72 + $3L$

Table 1. The four possible modes of operation and the bits required for each mode ($20ms$ frames). L is the number of subband energy values to encode

In Fig. 3, we plot the true excitation pattern and the excitation patterns corresponding to three of the modes. As is expected, the higher the bit-rate the closer the excitation patterns of the synthesized audio are to the true excitation.

2.2. Envelope Extension

The envelope extension algorithm is based on a constrained MMSE estimator that predicts the envelope based on features extracted from the lower band and energy values (loosely speaking) transmitted from the encoder (if necessary). The low band feature vector, denoted by $\mathbf{f}$, includes the narrowband cepstral coefficients, the pitch period, the spectral flatness, the spectral centroid and the zero crossing rate. Making use of these narrowband features and the transmitted energy values, we predict the high-band LPC cepstrum denoted by $\mathbf{y}$.

2.2.1. Formulating the Constrained MMSE Estimator

As stated earlier, the envelope extension algorithm has two different modes of operation. Assume that we are operating in modes 3 and 4 (the encoder transmits L energy values corresponding to L different subbands of the high band). The constrained MMSE estimator is encapsulated in equation (1):

$$\begin{aligned} &min_{\hat{\mathbf{y}}} && E[\|\mathbf{y}-\hat{\mathbf{y}}\|^2|\mathbf{f}] \\ &s.t. && 2\hat{\mathbf{y}}^T\mathbf{F}\mathbf{s}_1=\varepsilon_1, 2\hat{\mathbf{y}}^T\mathbf{F}\mathbf{s}_2=\varepsilon_2,\ldots,2\hat{\mathbf{y}}^T\mathbf{F}\mathbf{s}_L=\varepsilon_L \end{aligned} \quad (1)$$

where $\mathbf{s}_i$ is the selector vector corresponding to the i^{th} subband of the high band

$$\mathbf{s}_i^T = [00\ldots0\overbrace{11\ldots1}^{N/L}00\ldots0],$$

$\mathbf{F}_{i,j} = cos(i\frac{2\pi j}{N/2-1})$ $i = 0\ldots10$, $j = 0\ldots\frac{N}{2}-1$, N is the number of samples in the narrowband frame, and $2\hat{\mathbf{y}}^T\mathbf{F}\mathbf{s}_i$ represents the energy of the envelope in subband i in dB.

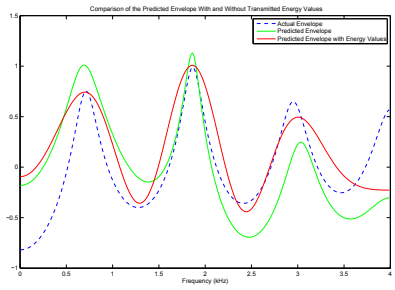

Fig. 4. The original high-band envelope, its MMSE estimate, and the constrained MMSE estimate

We can write the Lagrangian associated with (1) by writing a joint cost function that includes the function to be minimized and the constraints. The minimizer of this cost function will ensure that the energy of the envelope in certain bands is maintained at appropriate levels while also making use of the relationship between the extracted low-band features and the envelope of the missing band. It can be easily shown that the solution to (1) is given by (2):

$$\hat{\mathbf{y}} = \int \mathbf{y} p(\mathbf{y}|\mathbf{f}) d\mathbf{y} + \mathbf{F}\left(\lambda_1 \mathbf{s_1} + \ldots + \lambda_L \mathbf{s_L}\right), \qquad (2)$$

where the λ_i's can be computed from the constraints in (1).

2.2.2. Estimating the Joint Probability Distribution

In order to obtain a closed form solution for (2), it is necessary to estimate the multivariate probability distribution function that describes the joint statistical relationship between the input low-band features and the wideband envelope, $p(\mathbf{f}, \mathbf{y})$. A common practice for obtaining the probability distribution of large dimensional problems is to model the distribution using a weighted finite sum of Gaussians [17]. The joint distribution can then be written as follows:

$$p(\mathbf{f}, \mathbf{y}) = \sum_{k=1}^{K} a_k p_k(\mathbf{f}, \mathbf{y}) \qquad (3)$$

where $p_k(\mathbf{f}, \mathbf{y}) = N(\mathbf{C_k}, \boldsymbol{\mu_k})$. The parameters of this model, namely the $\mathbf{C_i}$'s and the $\boldsymbol{\mu_i}$'s, are estimated using the expectation maximization (EM) algorithm using approximately 30 minutes of training data obtained from the TIMIT database [18].

Given a pre-trained Gaussian mixture model (GMM) described by (3), we can derive the conditional distribution required to obtain the conditional expectation in (2). The parameters associated with each Gaussian in (3) can be decomposed into parts corresponding to $p_k(\mathbf{f})$ and $p_k(\mathbf{y})$ as shown below:

$$\mathbf{C_k} = \begin{bmatrix} \mathbf{C_k^{ff}} & \mathbf{C_k^{fy}} \\ \mathbf{C_k^{yf}} & \mathbf{C_k^{yy}} \end{bmatrix}, \boldsymbol{\mu_k} = \begin{bmatrix} \boldsymbol{\mu_k^f} \\ \boldsymbol{\mu_k^y} \end{bmatrix}. \qquad (4)$$

We can now write the conditional expectation in (2) in terms of the parameters of the pre-trained GMM as shown in (5).

$$\begin{aligned} \hat{\mathbf{y}} &= \sum_{k=1}^{K} a'_k \left(\boldsymbol{\mu_k^y} + \mathbf{C_k^{yx}} \mathbf{C_k^{xx}}^{-1} (\mathbf{f} - \boldsymbol{\mu_k^f})\right) \qquad (5) \\ &+ \mathbf{F}\left(\lambda_1 \mathbf{s_1} + \ldots + \lambda_L \mathbf{s_L}\right), \end{aligned}$$

where $a'_k = a_k \frac{p_k(\mathbf{f})}{\sum_{k=1}^{K} a_k p_k(\mathbf{f})}$.

The estimator in (5) may change depending on the mode of operation. If the rate determination algorithm deems the transmittal of extra information unnecessary, the estimator changes to an MMSE estimator without the additional correction factor to account for the transmitted energy values. For frames that require the extra information, pre-trained 3-bit VQ quantizers for the subband energy values are used for encoding. Four VQ quantizers are trained (1 for each subband of the high-band).

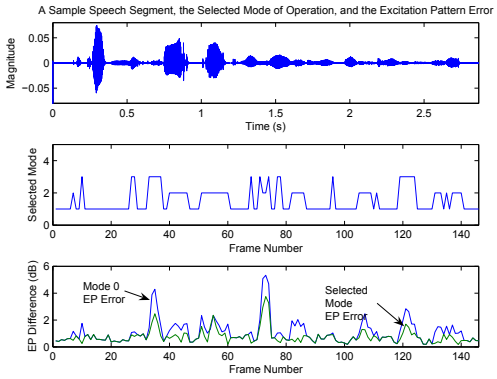

Fig. 5. Top: A sample speech segment. Middle: The selected mode of operation. Bottom: The excitation pattern errors for mode 1 and the selected mode

In Fig. 4 we show the true high-band envelope for a segment, the MMSE estimate of the envelope, and the constrained MMSE estimate of the envelope. As the figure shows, the constrained MMSE estimate is closer to the actual envelope than the envelope solely based on prediction.

2.3. Excitation Extension

The high-band excitation is estimated using one of two possible methods. For the modes of operation requiring no extra information (modes 1 and 3), the high-band excitation is simply a scaled and frequency translated version of the narrowband excitation. This simple method maintains the harmonic structure for voiced frames and the random-like structure for unvoiced frames. Mathematically, the excitation is given by:

$$u_{HB} = G \frac{w(n)}{\sqrt{\sum_{i=0}^{N} w(i)^2}} \sqrt{\sum_{i=0}^{N} u_{LB}(i)^2} \qquad (6)$$

where $w(n)$ is the translated version of the narrowband excitation, $u_{LB}(n)$ is the low band excitation, and G is the energy of the high-band excitation. G is estimated using the low band spectral tilt calculated using the approach in [11].

For the mode of operation requiring that the excitation be encoded, the algebraic codebook used in the adaptive multi-rate (AMR) algorithm is used [11]. The first three tracks are encoded using 3 bits, whereas the fourth is encoded using 4 bits. An additional 4 bits is required for the sign of each pulse. The full excitation is the sum of the fixed (algebraic) and the adaptive codebooks. The adaptive codebook gain is encoded using 4 bits whereas the delay (lag) is estimated from the lower band.

3. RESULTS

We evaluate results in terms of objective and subjective measures. The average log spectral distortion (LSD) over a number of speech segments is used to compare the proposed algorithm against a bandwidth extension algorithm that is solely based on prediction [16]. As Table 2 shows, the LSD associated with the proposed algorithm is smaller, indicating that the audio segments synthesized using the proposed approach are perceptually closer to the original. This is also shown in Fig. 5 in which we plot a sample speech segment (top); the mode that was selected by the rate determination algorithm on a frame- by- frame basis (middle); the difference in excitation patterns

(EP) between the coded signal and the original (bottom). The difference in the bottom plot is shown for both the selected mode and mode 1. There is a considerable reduction in the EP error when additional information is sent, thereby reducing the perceptible artifacts in the synthesized audio. This is also verified with subjective results below.

Database	Speech Sample	Prediction Based BWE	Proposed BWE algorithm
TIMIT	Male Speakers	4.03 dB	3.75 dB
TIMIT	Female Speakers	4.20 dB	3.62 dB

Table 2. A comparison of log spectral distortions between the proposed algorithm and an algorithm based only on prediction

In addition to the objective scores based on the LSD, informal listening tests were also conducted. In these tests we compare the proposed algorithm against the adaptive multi-rate narrowband encoder (AMR) operating at 12.2 kbps [11]. For the implementation of the proposed algorithm, we encode the low band (200 Hz - 3.4 kHz) of the signal at 7.95 kbps using AMR, and the high band (3.4 kHz - 7 kHz) using the techniques proposed in this paper. For all the experiments, a frame size of 20 ms was used. A group of 8 listeners was asked to state their preference between the proposed algorithm and the AMR 12.2 kbps for a number of different utterances. We compare the algorithms using utterances (not in the training set) from the TIMIT database [19] [18]. The results in Table 3 indicate that most listeners prefer the proposed algorithm in high SNR cases, however the results at low SNR scenarios are mixed. The reason for this is the introduction of the narrowband noise into the high band (through the extension of the excitation) becomes much more prominent in the low SNR scenario; therefore the speech is extended to the wideband, but so is the noise. On average, however, the results indicate that for approximately 3 kbps less we obtain clearer, more intelligible audio.

Database	Speech Sample	Preference Score	High Band Bit Rate
TIMIT	Male Speaker (clean)	75.0%	0.94 kbps
TIMIT	Female Speaker (clean)	87.5 %	0.91 kbps
TIMIT	Male Speaker (20 db SNR)	62.5 %	0.98 kbps
TIMIT	Female Speaker (20 db SNR)	87.5 %	1.12 kbps
TIMIT	Male Speaker (5 db SNR)	62.5 %	1.31 kbps
TIMIT	Female Speaker (5 db SNR)	50.0 %	1.22 kbps

Table 3. Preference scores for the proposed algorithm when compared to the AMR 12.2 kbps algorithm.

Database	Speech Sample	Artificial WB PS
CWP	Male Speaker	87.5 %
CWP	Female Speaker	87.5 %
TIMIT	Male Speaker (20 db SNR)	75.0 %
TIMIT	Female Speaker (20 db SNR)	62.5 %
TIMIT	Male Speaker (5 db SNR)	50.0 %
TIMIT	Female Speaker (5 db SNR)	37.5 %

Table 4. Preference scores for the algorithm operating in mode 1 (no extra information sent from encoder to decoder) under noisy conditions.

In addition to operating as a speech coder, the algorithm can be forced to operate in mode 1 (no extra information transmitted), thereby acting as a speech enhancement system. To examine this particular mode, we also include speech signals from the CWP database for the subjective tests. This corpus consists of noisy utterances from 336 cellular telephone users in a typical outdoor environment. The results are given in terms of preference scores (PS) again. Each listener is asked to compare the original noisy narrowband (NB) speech with the artificially extended wideband (WB) speech and select the preferred one. As the results show, for slightly noisy scenarios (utterances 1 - 4 in Table 4) the majority of listeners preferred the wideband signal citing improved "intelligibility" and a "crisper" sound. For the low SNR scenario (utterances 5 and 6 in Table 4), there is no clear preference in the results.

4. CONCLUSION

In this paper, we proposed a scalable speech coding method based on bandwidth extension. A psychoacoustic based rate determination algorithm selects an appropriate mode for each frame and decides on the number of extra bits required for a wideband representation. The envelope and excitation extension algorithms combine the transmitted information with predictive models in order to generate the final estimates. Results indicate that we can reduce the bit rate of a coded signal while improving quality. Although, the algorithm shows promise in its current form, further improvements are possible. Ongoing work focuses on better estimation methods for the excitation and improved performance in noisy conditions. In addition, quantization schemes specifically optimized for the problem at hand are also being studied.

5. REFERENCES

[1] A. Spanias, "Speech Coding: A tutorial review," in *Proceedings of the IEEE*. Vol 82, Issue 10. October, 1994.

[2] T. Unno and A. McCree, "A Robust Narrowband to Wideband Extension System Featuring Enhanced Codebook Mapping." *Proc. of ICASSP*. Philadelphia, PA.March, 2005.

[3] S. Yao and C. F. Chan, "Block-Based Bandwidth Extension of Narrowband Speech Signals by Using CDHMM." *Proc. of ICASSP*. Philadelphia, PA.March, 2005.

[4] C. F. Chan, W. K. Hui, "Wideband re-synthesis of narrowband CELP coded speech using multiband excitation model." *IEEE Proc. of ICSLP*. Vol 1, pp. 322-325. Philadelphia, PA. October, 1996.

[5] M. Nilsson and W.B. Kleijn "Avoiding over-estimation in bandwidth extension of telephony speech," *IEEE Proc of ICASSP*. Vol 2. 2001.

[6] P. Jax and P. Vary, "Enhancement of band-limited speech signals." *Proceedings of Aachen Symposium on Signal Theory*. pp. 331-336. Aachen, Germany. September, 2001.

[7] P. Jax and P. Vary, "An upper bound on the quality of artificial bandwidth extension of narrowband speech signals." *In Proc. of ICASSP*. vol 1, pp. 237-240. Orlando, FL.May, 2002.

[8] V. Berisha and A. Spanias, "Enhancing the quality of coded audio using perceptual criteria" in *IEEE Workshop on MMSP 2005*. October 2005.

[9] V. Berisha and A. Spanias, "Enhancing Vocoder Performance for Music Signals" in *Proc. of IEEE ISCAS*. May 2005.

[10] V. Berisha and A. Spanias, "Bandwidth Extension of Audio Based on Partial Loudness Criteria" in *Proc. of IEEE MMSP*. October 2006.

[11] 3GPP TS 26.190, *AMR-WB; Transcoding Functions*. 2001.

[12] V. Atti and A. Spanias. "Rate determination based on perceptual loudness" in *IEEE Proc. of ISCAS 2005*. Vol 2, pp. 848-851. May 2005.

[13] B.C.J. Moore, *An Introduction to the Psychology of Hearing*. Fourth Edition, 1997.

[14] T. Painter and A. Spanias, "Perceptual Segmentation and Component Selection for Sinusoidal Representations of Audio", *IEEE Trans. on Speech and Audio Processing*. Vol 13, pp. 149-162, 2005.

[15] M. Dietz, L. Liljeryd , K. Kjorling, and O. Kunz, "Spectral band replication, a novelapproach on audio coding." *Proc. of AES*, 2002.

[16] J. Epps. "Wideband Extension of Narrowband Speech for Enhancement and Coding," *PhD Thesis*, The University of New South Wales, 2000.

[17] G. McLachlan and D. Peel "Finite Mixture Models," *Wiley*, 2000.

[18] J. S. Garofolo, Lori F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. "The DARPA TIMIT acoustic-phonetic continuous speech corpus", February 1993.

[19] R. A. Cole, M. Fanty, M. Noel, and T. Lander. "Telephone speech corpus development at CSLU." *In Proc. of ICSLP*, Yokohama, Japan, 1994.

[20] Y. M. Cheng et al., "Statistical Recovery of Wideband Speech from Narrowband Speech", *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 544–548, Oct. 1994.