Speech Bandwidth Extension by Improved Codebook Mapping Towards Increased Phonetic Classification

Rongqiang Hu, Venkatesh Krishnan, David V. Anderson

Center for Signal and Image Processing
Georgia Institute of Technology, Atlanta, GA, USA
(rqhu, vkrish, dva)@ece.gatech.edu

Abstract

Bandwidth limitation (0-4KHz) is a major degradation for the performance of the current speech communication systems. The narrowband speech provides much lower quality and intelligibility than wideband speech (0-8KHz). Speech bandwidth extension technology has been recently investigated to aim at artificially regenerating the missing high-band speech signal. This paper describes a robust speech bandwidth extension system by an improved codebook mapping method (CM-IPC), which includes a modified codebook training towards increased phonetic classification, marginal LSF interpolation, codebook mapping with memory, and codebook interpolation. A variety of experiments, in clean and noisy environments, were conducted to evaluate the performance of the proposed system. The results indicate the improvement in objective quality measures. The proposed system can also be applied to the feature extension estimation of speech signals.

1. Introduction

In many speech transmission systems, such as the digital public telephone system, low-bit-rate-speech coding environment (MELP), the bandwidth of speech is limited to 4KHz. This kind of speech is so called "telephone speech". Compared to natural speech, telephone speech has a significantly degraded performance. The bandwidth limitation of telephone speech reduces speech intelligibility by about 10 percent, and decreases the subjective quality score, which is measure in terms of the subjective mean opinion score (MOS) by more than one point [1].

Owing to the importance of the acoustic bandwidth for speech intelligibility and especially for subjective quality. It is worthwhile to extend the speech bandwidth. Particularly, in digital communication and hands-free telephony, there is a demand for enhancing the subjective speech quality.

To avoid the modification of narrowband communication systems, where the receiver does not have the access to wideband signals, recent work [2, 3, 4, 5, 6] has been done to artificially extend the narrowband speech to wideband speech by signal processing techniques. These techniques are motivated from the fact that the spectral envelope of the lower and upper frequency bands of the speech signal are dependent, which can be illuminated from the speech production model. Within the state-of-the-art speech bandwidth extension techniques, codebook mapping is a commonly used and promising method.

In this paper, we present a speech bandwidth extension system using improved codebook mapping towards increased phonetic classification based on $aprior$ knowledge.

2. System Overview

The proposed speech bandwidth extension system, shown in Figure 1, consists of two major modules: excitation extension and spectral envelope extension. The received narrowband speech signal is first analyzed. The narrowband spectral envelope representatives, line spectral frequencies (LSFs), are obtained. The residual of the analyzer is the narrowband excitation. Then, the excitation and spectral envelope of the narrowband are extended to wideband using corresponding extension models. In the synthesis part, The extended wideband excitation and spectral envelope are synthesized. The output is passed through a high-pass filter, and is summed with the narrowband speech, which is upsampled by a factor of 2 and passed through a low-pass filter.

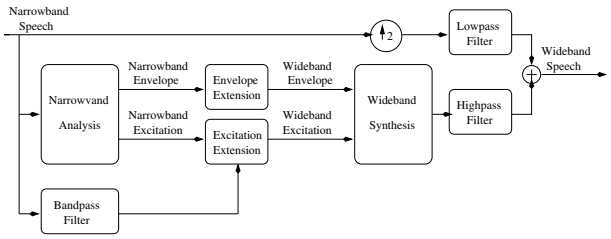


Figure 1: Block diagram of the proposed system

2.1. Excitation Extension

In wideband coding technology, pitch adaptive modulation has been shown to provide better wideband excitation. This was confirmed in speech bandwidth extension in [7]. The performance is better than spectral folding and spectral replication. In our excitation extension module, we use the envelope of the bandpass signal between 2.5-3.4KHz to modulate the wideband excitation.

2.2. Spectral Envelope Extension

A number of techniques for estimating wideband spectral envelope have been recently proposed [2, 3, 4, 5, 6]. In particular, codebook mapping is popular used and promising. The approach is based on a pair of coupled codebooks that contain representations of the spectral envelopes of the narrowband and wideband speech, respectively. For each signal frame, the spectral envelope of the narrowband speech signal, represented by the feature vector x is compared to a list of typical narrowband spectral envelopes that are stored in a pre-trained primary codebook. The closet codebook entry is selected. In parallel to

the searched primary codebook, there exists a second codebook, which contains corresponding wideband spectral envelope representatives. The estimate $\hat{y}$ of the wideband spectral envelope is the entry of the second codebook that is assigned to the selected entry of the primary codebook.

There are many representatives for spectral envelope, including linear prediction coefficients (LPC), Mel-frequency cepstral coefficients (MFCC), LSF and so on. In this application, the spectral envelope is stored as LSF. LSF coefficients describe the spectral envelope in a more direct way. The range of (0, π) corresponds proportionally to the whole frequency range of the spectrum. Additionally, LSF has some desirable properties: when LSF values fall in the range (0, π), the recovered LPC filter has guaranteed stability; local errors of LSF values only cause local spectral distortion. Therefore, codebook mapping based on LSF values is more tolerant to estimation errors, as a single error cannot harm the whole spectral envelope. A comparison between LPC and LSF has been made in [6]. More important, the selection of LSF makes it easier to implement an improved codebook mapping method, distance measures based on $aprior$ knowledge. This will be covered in later section.

2.3. Assessment

It is well-known that the synthesized speech by bandwidth extension provides higher intelligibility, especially for the fricatives where the most spectral energy locates in high-band, and better quality also. Log spectral distortion (LSD) is an useful for assessing the performance. The LSD in overall bands (narrowband and high-band) indicates the improvement of a bandwidth extension relative to narrowband speech. To provide a comparison between different bandwidth extension methods, the high-band LSD (HB-LSD) is used, since the actual narrowband signal is given.

The performance of the speech bandwidth extension using codebook mapping depends on many factors, in which codebook size and codebook partition method (distance measure) are crucial. Recent research has found that the codebook mapping performance in the speech bandwidth extension saturates for codebook sizes greater than about 256 [5]. Therefore, 256-level codebooks are used in our system. Both narrowband and high-band parameters were 12th-order LSF in the implementation of conventional codebook mapping.

In the preliminary experiments, a speech bandwidth extension system using conventional codebook mapping is implemented. The overall LSD between the synthesized wideband speech and original narrowband speech is indicated in Table 6.

Table 1: The performance of speech bandwidth extension using conventional codebook mapping

	Narrowband	BWE
Overall LSD (dB)	11.45	**5.32**

3. Improved Codebook Mapping

In the proposed system, we apply three techniques to improve the performance of codebook mapping: a distance measure towards increased phonetic classification, marginal LSF interpolation, codebook mapping with memory, and codebook interpolation.

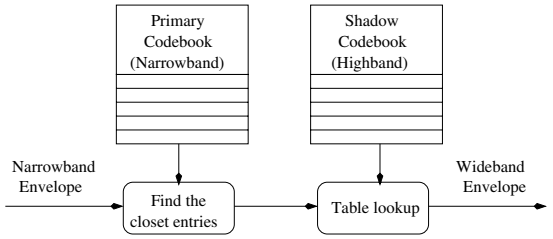


Figure 2: Block diagram of the conventional codebook mapping method

3.1. Codebook Training Towards Increased Phonetic Classification

The objective for codebook mapping in bandwidth extension is to minimize the high-band spectral distortion given the narrowband spectral representatives. The criterion is to minimize the mean square error of the LSF parameters in high-band.

$$min[d(k)] = min\{E[\|y(k) - \hat{y}(k)\|^2 \,|x(k)\,]\} \qquad (1)$$

Where x and y are the LSF parameters in narrow-band and high-band.

In conventional codebook mapping scheme, the minimization is implemented in two steps. First, the entries of the primary codebook C_x are defined by the vector quantization (VQ) of the speech feature $x(k)$. The quantization mapping Q is defined such as to minimize the criterion function $d(x(k), \hat{x}_i(k))$ between the input vector $x(k)$ and the entry $\hat{x}_i(k)$,

$$Q_s(x(k)) = \underset{\hat{x}_i(k) \in C_{x(k)}}{\arg\min} \; d(x(k), \hat{x}_i(k)) \qquad (2)$$

where the distance measure $d(x(k), \hat{x}_i(k))$ is defined as Euclidian distance

$$d(x, \hat{x}_i) = \|x - \hat{x}_i\|^2 \qquad (3)$$

Then, the corresponding entry in the shadow codebook is defined as:

$$\hat{y}_i = \underset{\hat{y}}{\arg\min}\, E\{d(y, \hat{y}_i) | x \in \gamma_i\} \qquad (4)$$

where γ_i is the quantizer cell assigned to code vector $\hat{x}_i$. This equation is solved by the conditional expectation $\hat{y}_i = E\{y|x \in \gamma_i\}$. The expectation can be determined using a large number of pairs of training vectors $\{x(m), y(m)\}, m = 1...N_m$ by averaging the vectors $y(m)$ extracted from those signal frames for which $x(m) \in \gamma_i$.

It is well-known that each phoneme class has distinctive acoustic properties, thus distinctive LSF parameters in narrowband and high-band. It means that better estimation of high-band LSF parameters can be achieved if the phoneme class of the signal in current frame is given. Therefore, equation (1) can be refined as:

$$min[d(k)] = min\{E[\|y(k) - \hat{y}(k)\|^2 \,|x(k), x(k) \in \vartheta(k)\,]\} \qquad (5)$$

Where $\vartheta(k)$ represents the phonetic label in the kth frame.

Generally, the codebook mapping and segmentation algorithm are computation complex. So, it is not practical to incorporate a segmentation algorithm in bandwidth extension. The other problems of using the segmentation are the variation of the split codebooks and large knowledge to be trained. The performance highly depends on the phonetic classification rate. Many artifacts will be introduced due to the misclassification.

Therefore, in the proposed algorithm, we modify the conventional codebook mapping scheme towards increased phonetic classification. The entry in the shadow codebook is alternatively defined as:

$$\hat{y}_i = \arg\min_{\hat{y}_i} E\{d(y, \hat{y}_i)|x \in \vartheta_i\} \quad (6)$$

In this case, the training of the primary codebook is optimized to the following criterions:

$$Q(x) = \arg\min_{\hat{x}_i \in C_x} d(x, \hat{x}_i) \quad (7)$$

$$Q(x) = \arg\max_{\hat{x}_i \in C_x} P(\hat{x}_i \in \vartheta_i | x \in \vartheta_i, x \in \gamma_i) \quad (8)$$

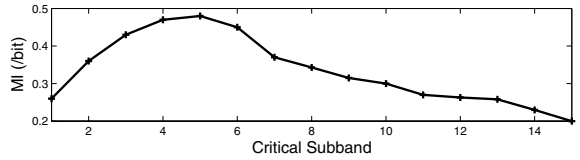
There is no explicit method to optimize both criterions. In [8], the mutual information (/bit) between the short-time critical-band logarithm spectral energy and phonetic classification was found. The measure is for narrow-band signal, as shown in figure 3. Since a shift made on one LSF parameter is only related to local spectral distortion, thus change the logarithm spectral energy in the corresponding critical-band. This information can be employed in the training of the primary codebook training, in order to increase the phonetic classification. A simplified criterion is used in the our system,

$$Q(x) = \arg\min_{\hat{x}_i \in C_x} \hat{d}(x, \hat{x}_i) \quad (9)$$

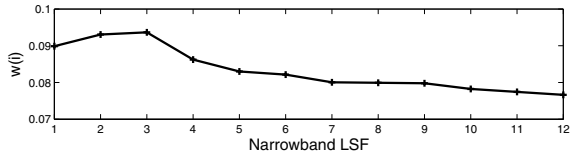
where the distance function is defined as:

$$\hat{d}(x, \hat{x}_i) = \sum_{n=1}^{N} w(n) \cdot (x(n) - \hat{x}_i(n))^2 \quad (10)$$

The parameter N is the dimension of the vector x, weights w are illuminated in figure 3.



(a)



(b)

Figure 3: (a) The mutual information (/bit) between the short-time critical-band logarithm spectral energy and phonetic classification, (b) The proposed weighting for a distance measure toward increased phonetic classification.

To evaluate the proposed distance measure, we define the hit rate of phonetic classification of a quantized cell γ_i as the maximal phonetic probability associated.

$$P(\gamma_i) = \arg\max_{\vartheta_i} P(\hat{x}_i \in \vartheta_i | x \in \vartheta_i, x \in \gamma_i) \quad (11)$$

Table 2: The hit rate of phonetic classification based on codebook mapping

	Conventional	IPC
vowel	64.23%	**67.19**%
unvoiced fricative	54.47%	**61.01**%
nasal	62.73%	**70.44**%

Average hit rates in the overall primary codebook are measured and shown in table 2.

The phonetic classification hit rates are increased by the codebook partition using the proposed distance function. The improved HB-LSD is also achieved, as indicated in table 3.

Table 3: The performance of the improved codebook mapping towards increased phonetic classification (IPC)

	Conventional	IPC
HB-LSD (dB)	8.01	**7.12**

3.2. Marginal LSF interpolation

This algorithm makes the assumption that the frequency ranges $(0, \pi/2)$ and $(\pi/2, \pi)$ have the same number of LSF values in wideband speech. For example, when the LPC order is 24, there are always 12 LSFs in $(0, \pi/2)$ and 12 LSFs in $(\pi/2, \pi)$. This assumption is not true for all frames, and the actual distribution is (11,13) or (13,11) with a probability of around 50% in the training by TIMIT database. Particularly, the high-frequency consonants, which are of special interest to our problem, have mainly the distribution of (11,13). This reflects the fact that the speech energy is concentrated in the high-band. To compensate for this drawback, the proposed algorithm uses marginal LSF interpolation. In the shadow codebook, the upper 13th-order LSF (12-24) is used instead of the upper 12th-order LSF (13-24). The mean of the lowest LSF in shadow codebook and the highest LSF in primary codebook will replace the highest LSF of actual narrowband speech in synthesis. The method is effective to lower the spectral distortion in the frequency region of (3KHz-5KHz).

Table 4: The performance of the marginal LSF interpolation

	Non-overlap codebook	Marg. LSF interp.
HB-LSD (dB)	7.12	**7.01**

3.3. Codebook Mapping With Memory

The conventional codebook mapping scheme is memoryless. The estimation of high-band LSF is based on the current time frame only. From the speech production model, there is correlation between each frame. Therefore, utilizing time history information can enable a more accurate codebook mapping.

$$min\{E[\|y(k) - \hat{y}(k)\|^2 | x(k), x(k-1), \hat{y}(k-1), x(k) \in \vartheta(k)]\} \quad (12)$$

We update the estimation of high-band LSF based on the interpolation of previous high-band LSF using the distance between their narrowband LSFs.

$$\hat{y}(k) = C_{ipc}[x(k)] + F[\hat{d}(x(k), x(k-1))] \cdot \hat{y}(k-1) \quad (13)$$

where, $C_{ipc}[\cdot]$ denotes the codebook mapping as described above, $F[\cdot]$ is a smoothing function by experiments.

Table 5: The performance of codebook mapping with memory

	Memoryless	With memory
HB-LSD (dB)	7.01	**6.79**

3.4. Codebook Interpolation

It is shown, instead of a simple table lookup, the estimate $\hat{y}(k)$ is determined by a weighted sum of the most probable codebook entries.

$$\hat{y}(k) = \sum_{m=1}^{M} \alpha_m \hat{y}_m(k) \quad (14)$$

The individual weights α_m are inverse portional to the distance of the narrowband feature vector $x(k)$ to the respective mth closed primary codebook entry $\hat{x}_m(k)$.

Table 6: The performance of codebook mapping using interpolation

	NO interpolation	interpolated
HB-LSD (dB)	6.79	**6.63**

4. Evaluation in Noisy Environment

Speech bandwidth extension is challenging in noisy environments, because a speech bandwidth extension system requires accurate estimation of narrowband spectral parameters. If the narrowband speech is corrupted by background noise, the estimation error in narrowband will introduce large error in the estimation in high-band. In this paper, besides the performance evaluation in clean condition, we conducted experiments on noisy speech. The evaluation is therefore important to indicate the real-world application of a speech bandwidth extension system, which has not been measured yet.

Our experiments were conducted in various additive noise conditions. The level of noise degradation is characterized by the narrowband LSD (NB-LSD). Two common noise types are used: room noise and car noise. The noisy signal is first processed by a noise suppressor. Then, the enhanced narrow speech is used for extending speech bandwidth. The overall LSD relative to clean speech is measured to indicate the performance improvement, as shown in Table 7, where "CM" denotes the conventional codebook mapping method.

5. Conclusion

We proposed a speech bandwidth extension system by improved codebook mapping towards phonetic classification (CM-IPC). The model is simplified using a weighted distance function based on the mutual information between frequency bands and phonetic labeling. A set of techniques are used to improved the performance by marginal LSF interpolation, codebook mapping with memory, and codebook interpolation. The proposed system is effective in reducing the spectral distortion, thus help to increase the objective quality of speech.

Table 7: The performance of the speech bandwidth extension in noisy conditions

Noise Type	NB-LSD (dB)	Overall LSD (dB)		
		Noisy	CM	CM-IPC
Room	1.12	11.87	6.48	**5.43**
	2.06	12.35	7.07	**5.89**
	3.07	13.41	7.91	**6.72**
Car	1.07	11.23	6.28	**5.40**
	2.01	12.02	6.94	**5.87**
	3.11	13.35	7.87	**6.65**

6. References

[1] E. Larsen and R. M. Aarts, *Audio Bandwidth Extension: Application of Pstchoacoustic, Signal Processing and Loudspeaker Design*. Whitley, 2004.

[2] J. Epps and W. Holmes, "A new technique for wideband enhancement of coded narrowband speech," in *Proceedings of the IEEE Workshop on Speech Coding*, Porvoo, Finland, Sept. 1999, pp. 174–176.

[3] D. Heide and G. Kang, "Speech enhancement for bandlimited speech," in *Proceedings of ICASSP*, vol. 1, Seattle, USA, May 1998, pp. 393–396.

[4] N. Enborn and W. B. Kleijn, "Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients," in *Proceedings of the IEEE Workshop on Speech Coding*, Porvoo, Finland, Sept. 1999, pp. 171–173.

[5] P. Jax, *Enhancement of Bandlimited Speech Signals: Algorithms ans Theretical Bounds*. Ph.D Thesis, RWTH Aachen, 2002.

[6] T. Unno and A. McCree, "A robust narrowband to wideband extension system featuring enhanced codebook mapping," *Proceedings of ICASSP*, Mar. 2005.

[7] Y. Qian and P. Kabal, "Dual-mode wideband speech recovery from narrowband speech," in *Proceedings of EUROSPEECH*, Sept. 2003, pp. 1433–1437.

[8] H. H. Yang, S. Sharma, S. van Vuuren, and H. Hermansky, "Relevance of timefrequency features for phonetic and speakerchannel classification," *Speech Communication*, vol. 31, pp. 35–50, Aug. 2000.