Objective Analysis of the Effect of Memory Inclusion on Bandwidth Extension of Narrowband Speech

Amr H. Nour-Eldin, Peter Kabal

Department of Electrical & Computer Engineering
McGill University, Montréal, Québec, Canada
amr.nour-eldin@mail.mcgill.ca, peter.kabal@mcgill.ca

Abstract

For the purpose of improving Bandwidth Extension (BWE) of narrowband speech, we continue our recent work on the positive effect of exploiting the temporal correlation of speech on the dependence between speech frequency bands. We have shown that such memory inclusion into MFCC speech parametrization translates into higher *highband certainty*. In the work presented herein, we employ VQ to estimate highband discrete entropies, thus refining our analysis of the effect of memory inclusion on increasing highband certainty. Moreover, we extend our previous analysis to LSF parameters. We further construct a BWE system that exploits our memory inclusion technique, thus translating highband certainty gains into practical BWE performance improvement as measured by the objective quality of reconstructed speech. Results show that memory inclusion decreases the log-*Spectral Distortion* of the extended highband speech by as much as 1 dB corresponding to more than 14% relative.
Index Terms: Bandwidth Extension, Mutual Information.

1. Introduction

In traditional telephone networks, speech bandwidth is limited to the 0.3–3.4 kHz range. As a result, narrowband speech has sound quality inferior to its wideband counterpart and it shows reduced intelligibility especially for consonant sounds. Wideband speech reconstruction through Bandwidth Extension (BWE) attempts to regenerate the low (20–300 Hz) and highband (3.4–7 kHz) signals lost during the filtering processes employed in traditional networks, thereby providing backward compatibility with existing networks.

BWE is based on the assumption that narrowband speech correlates closely with the highband signal, and thus, given some a priori information about the nature of this correlation, the higher frequency speech content can be estimated. Although significant research has been published on BWE techniques, few researchers have investigated the correlation assumption between the narrow and highband spectral envelopes. In [1], a rough lower bound on the Mutual Information (MI) between narrow and high frequency bands was derived. This initial attempt, however, did not present a meaningful conclusion in terms of BWE. This work was extended in [2] to quantify the certainty about the high band given the narrow band by determining the ratio of the MI between the two bands to the discrete entropy of the high band. The authors show that this ratio (representing the dependence between the two bands) is quite low. Despite this low dependence, BWE schemes have continued to use *memoryless mapping* between spectra of both bands.

In our recent work [3], we exploited the considerable temporal correlation properties of speech by including memory in Mel Frequency Cepstral Coefficient (MFCC) speech parametrization (through *delta features*). These features are obtained through linearly weighted differences between neighbouring conventional *static* feature vectors. Similar to [2] and [4], MI and highband entropy are estimated using the numerical method of stochastic integration, where the marginal and joint distributions of the narrow and high band parameterizations are modelled by Gaussian mixture models (GMMs) for both static and extended (static+delta) acoustic spaces. Our results showed that such memory inclusion into speech parametrization translates into higher highband certainty (as measured by the ratio of MI to discrete high band entropy). Replacing half of the static features (per speech vector) by delta coefficients results in a 216% relative increase of highband certainty. By varying the widths of the window of static feature vectors involved in the estimation of the delta features, we demonstrated that the greatest gains in highband certainty were those corresponding to short-term memory inclusion ($t \lesssim 160$ msec), representing roughly inter-phoneme temporal information. Phonemes with mostly highband energy, e.g., fricatives, stand to have the most benefit of such short-term triphone memory inclusion. Since BWE schemes generally perform poorly when reconstructing such phonemes, we concluded that the performance of such BWE schemes is expected to be considerably improved by triphone-specific memory inclusion.

We continue this work by extending the analysis to Line Spectral Frequencies (LSFs). LSFs are widely used in speech coding, and are particularly attractive for BWE for their quantization error resilience and perceptual significance properties. Furthermore, we improve the accuracy of our discrete highband entropy estimates (used to estimate highband certainty) through vector quantization (VQ) of the highband feature vector space, rather than using the scalar quantization approximation of [2].

Finally, we translate the highband certainty gains obtained through memory inclusion into practical BWE performance improvement by incorporating our technique in an LP-based *dual-mode* BWE system [5]. Objective analysis of the reconstructed speech quality shows that memory inclusion decreases the log-*Spectral Distortion* (SD) of the extended highband speech (versus that obtained by BWE with conventional static features) by as much as 1 dB corresponding to $> 14\%$ relative.

2. Information measure estimation

Representing the narrow and high bands by the continuous (vector) variables X and Y, respectively, the mutual information can be written in terms of the joint and marginal *pdf*s as

$$I(X;Y) = \int_{\Omega_Y}\int_{\Omega_X} f_{XY}(x,y) \log_2 \left(\frac{f_{XY}(x,y)}{f_X(x) f_Y(y)} \right) \mathrm{d}x \, \mathrm{d}y. \quad (1)$$

Following [2] and [4], we model the densities using the GMM $f_{GMM}(x,y) = \sum_{m=1}^{M} \alpha_m f_G(x,y|\theta_m)$, where M is the number of mixture components, α_m is the mth mixture weight and $f_G(\cdot)$ denotes the mutlivariate Gaussian distribution defined by the mean vector and covariance matrix in $\theta_m = \{\mu_m, C_m\}$. Rewriting Eq. (1) as $I(X;Y) = \mathrm{E}\left[\log_2\left(\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}\right)\right]$ and replacing the expectation operator by the sample mean yields (for N samples) $I(X;Y) \approx \frac{1}{N}\sum_{n=1}^{N}\log_2\left(\frac{f_{XY}(x_n,y_n)}{f_X(x_n)f_Y(y_n)}\right)$. Thus, MI can be estimated (in bits) using numerical integration by substituting the *pdf*s for their GMM estimates; i.e.,

$$\hat{I}(X;Y) = \frac{1}{N}\sum_{n=1}^{N}\log_2\left(\frac{f_{GMM}(x_n,y_n)}{f_{GMM}(x_n)f_{GMM}(y_n)}\right). \quad (2)$$

Similarly, an estimate of the differential entropy of Y can be obtained by $\hat{h}(Y) = -\frac{1}{N}\sum_{n=1}^{N}\log_2\left(f_{GMM}(y_n)\right)$. Since $\hat{h}(Y)$ is susceptible to variable scaling, the discrete entropy $H(Y)$ provides a more consistent estimate of highband self-information. Following [2], we approximated $H(Y)$ in [3] by $H(Y) \approx \hat{h}(Y) - \log_2(\Delta^K)$, where K is the dimension of the vector Y. This approximation is only valid if the quantization step-size Δ is small enough such that the *pdf* of Y can be considered flat in each quantization bin (high-rate assumption). Moreover, since this approximation implies scalar quantization of the continuous space of Y, a certain distortion is introduced which increases with the dimensionality, K, of Y (in [3] and in this work, we use a maximum K=6).

By rather employing VQ of the Y space, we obtain more accurate $H(Y)$ estimates through VQ's space filling, shape, and memory advantages over scalar quantization [6]. Using a codebook size of 256, we perform VQ using the Lloyd training algorithm [7] (k-means clustering) with Euclidean distances as the distortion measure. This results in a codebook with minimum mean distortion estimated over all 256 cells for a training-set of 10^5 MFCC/LSF feature vectors obtained from the TIMIT speech corpus [8]. Accordingly, if i is the VQ cell index and V_i is the ith cell, then the discrete highband entropy is estimated by

$$H(Y) = -\sum_{i} P_{Y_i}(y)\log_2\left(P_{Y_i}(y)\right), \quad (3)$$

where $P_{Y_i}(y) = P(Y = y \in V_i) = \frac{N(y \in V_i)}{10^5}$.

3. Speech parametrization and modelling

In our previous work, [3], we chose MFCCs to parameterize the narrow and high bands since they can be directly related to SD [9]; an objective speech quality measure widely used to assess the performance of spectral envelope quantizers. Furthermore, as MFCCs are calculated using a DCT, they have the desirable property of being decorrelated for different speech classes. MFCCs were shown to provide the highest class separability among most common spectral envelope parameters [10]. The advantage in terms of our work is that employing MFCCs results in GMMs with higher discriminative ability between different speech classes, which in turn results in more accurate modelling of the acoustic space, and hence, better MI estimates.

LSFs are particularly attractive for BWE systems due to their properties of robustness against quantization noise and perceptual significance. Accordingly, we extend our analysis of the effect of memory inclusion on highband certainty to LSF parameters. Although LSFs are less discriminative of speech classes than MFCCs, their resilience to quantization noise render our $H(Y)$ estimates more accurate than those obtained using MFCCs. Moreover, the perceptual significance of LSFs (where properties of formants and valleys can be related to LSF pairs) implies the improved ability of GMMs to capture perceptually significant characteristics of speech through modelling the (LSF-parameterized) acoustic space. More importantly, LSFs have the important advantage of being easily convertible into linear prediction (LP) coefficients, and hence, coupled with an excitation estimate, can be directly used for BWE. In addition, SD can be straightforwardly estimated from LP-coefficients.

We explicitly capture speech temporal information by including memory directly in the spectral envelope parametrization in the form of *delta* coefficients appended to the MFCC/LSF vectors. Delta coefficients are obtained from the *static* coefficient vectors by a first-order regression implemented through linearly weighted differences between neighbouring static vectors. Since immediately successive frames show only minor differences between their parameters, the trajectory of parameter variation with time is more accurately and easily identified as the time separation between the involved static frames increases. Hence, the difference weights increase in proportion to the distance (in frames) between the two static vectors whose difference is being evaluated. Delta coefficients are calculated via:

$$\delta_t = \frac{\sum_{\theta=1}^{\Theta}\theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta}\theta^2}, \quad (4)$$

where δ_t is a delta coefficient vector at time t computed from the corresponding static coefficient vectors $c_{t-\theta}$ to $c_{t+\theta}$, and Θ specifies the number of neighbouring static frames to consider.

Speech parametrization is detailed in [3] for MFCCs. LSF parametrization follows in a similar manner where narrowband and highband speech are obtained through bandpass (0.3–3.4 kHz) and highpass (3.7–8 kHz) filtering of the TIMIT wideband speech, followed by linear prediction and conversion to LSFs. Delta coefficients are then calculated by Eq. (4) and appended to the static vectors for the extended space case. As the ratio of highband to narrowband energy represents an important measure of dependence between both bands, frame log-energy (and its delta coefficient) was also appended to feature vectors of each band in the static (and extended) space case(s).

GMM modelling of the MFCC/LSF spaces is performed as described in [3], where a single GMM, with 107 full covariance mixtures, is used to model the approximately 40 English phonemes. To increase the amount of available data, 20 msec frames with 50% overlap were extracted form the 3696 training and 1344 test speech files available in the TIMIT database, resulting in 1,126,746 training and 411,620 test frames.

4. Effect of memory inclusion

Our analysis in [3] led us to conclude that narrowband delta features contain no information about the static high band, and hence, replacing or appending narrowband static coefficients by their delta ones adds no benefits. In contrast, including memory in both narrow and high bands results in significant gains in both MI and highband certainty. Accordingly, the analysis that follows considers the latter case only. The extent of memory included (represented by the number of neighbouring static vectors, Θ, involved in the estimation of the delta coefficients) was also shown to have a considerable impact on the increase in highband certainty. The greatest gains in highband certainty were those corresponding to short-term memory inclu-

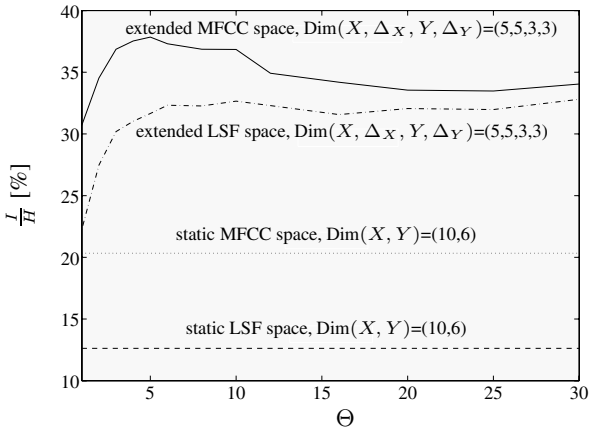

Fig. 1: *Ratio of MI to highband entropy ($\frac{I}{H}$) versus Θ for the MFCC and LSF extended spaces. The $\frac{I}{H}$ ratios for the static spaces with Dim$(X, Y) = (10, 6)$ are also shown for reference.*

	Dim(X, Y)	$I(X;Y)$	$H(Y)$	$\frac{I(X;Y)}{H(Y)}$	β_{max}
MFCCs	(10,6)	1.59	7.82	20.3%	17.5%
	(5,3)	1.48	7.87	18.8%	19.1%
LSFs	(10,6)	0.84	6.67	12.6%	20.2%
	(5,3)	1.18	7.93	14.9%	17.9%

Table 1: *Information measures (in bits) and highband certainty results for the MFCC and LSF static spaces.*

sion ($\Theta \lesssim 8 \equiv t \lesssim 160$ msec), representing roughly inter-phoneme temporal information.

Let X and Y represent the static MFCC/LSF vectors of the narrow and high bands, respectively, with Δ_X and Δ_Y representing the corresponding delta coefficient vectors. Then, the increase in certainty about the high band is given by

$$\beta \triangleq \frac{I(X, \Delta_X; Y, \Delta_Y)}{H(Y, \Delta_Y)} - \frac{I(X;Y)}{H(Y)}, \qquad (5)$$

where the first fraction represents the highband certainty for the extended space, and the second that of the static space.

Using GMMs to estimate $I(\cdot;\cdot)$ per Eq. (2), and VQ of the highband feature vectors to estimate $H(\cdot)$ per Eq. (3), we obtain the highband certainty results illustrated in Fig. 1 for varying widths, Θ, of the time window used to calculate delta features. Fig. 1 shows the estimated highband certainty for the extended MFCC/LSF spaces, with Dim$(X, \Delta_X, Y, \Delta_Y)$=(5,5,3,3), versus two reference static spaces with Dim(X, Y)=(10,6). The extended space are, thus, obtained by replacing half the static features by the corresponding delta ones. Table 1 additionally lists the results obtained for the case where the extended spaces are obtained by appending the static spaces. Table 1 also lists the maximum increase in highband certainty, β_{max}.[1]

Using β_{max} and the highband certainty results for the static spaces of Table 1, the *relative* increase in highband certainty due to memory inclusion reaches a maximum of 102% for the static (5,3) MFCC space, and 160% for the (10,6) LSF space. As we concluded in [3], this clearly demonstrates the strong effect of memory inclusion on increasing highband certainty, and consequently, BWE performance. Fig. 1 further illustrates this effect.

For the static spaces considered in Table 1, both MI and highband certainty figures show that MFCC parameters outperform LSFs in terms of capturing information mutual to both narrow and high bands. This observation is further confirmed by the fact that both MI and highband certainty increase with increasing MFCC dimensionality, in contrast to their decrease using LSFs. Furthermore, while the absolute gains in highband certainty due to memory inclusion are equivalent for both sets of parameters, Fig. 1 shows the superiority of MFCCs over LSFs in terms of overall highband certainty for the same dimensionality. Notwithstanding the advantages of LSFs over MFCCs; namely quantization noise robustness and more accurate and straightforward speech reconstruction, the observations above lead us to conclude that, in principle, BWE based on MFCC parametrization with memory inclusion is inherently better.

[1] The highband certainty results of our previous work in [3] deviate from those of Table 1 above due to the afore-mentioned distortion introduced by the scalar quantization approximation used for the estimation of the discrete higbhand entropy, $H(Y)$, particularly for increasing K.

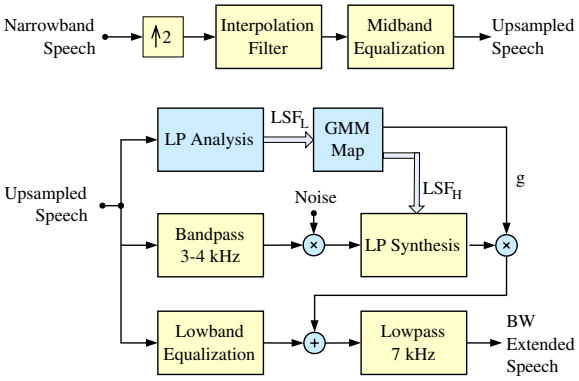

Fig. 2: *The dual-mode bandwidth extension system.*

5. BWE with memory inclusion

To evaluate the effect of memory inclusion on BWE performance, we employ a *dual-model* BWE system based on that of [5] shown in Fig. 2. This system exploits equalization to expand the apparent bandwidth of narrowband speech. Equalization is applied both at low frequencies as well as at high frequencies to push the bandwidth out to 100 Hz at the low end and up to 4 kHz at the high end. The equalization algorithm is more accurate than any estimation algorithm can be in this frequency range. Furthermore, as an additional benefit, the equalized signal is bandpass-filtered (3–4 kHz range), followed by Gaussian noise modulation, to produce an enhanced excitation signal which is used for signal reconstruction in the region above 4 kHz. GMM statistical estimation is used to generate the complementary spectrum, represented by LSFs, in the range from 4 to 7 kHz. The estimated highband LSFs, converted to LP-coefficients, are then used together with the estimated excitation signal to reconstruct the highband speech through LP synthesis.

In addition, an excitation gain, g, is used to scale the synthesized highband components such that the energy of the reconstructed highband components is equal to that of the corresponding frequency band in the orginal wideband speech used for GMM training. The excitation gain is calculated as the square root of the energy ratio of the original highband signal to the resynthesized one. As g is assumed to be correlated with the narrowband spectrum, it can be statistically estimated from

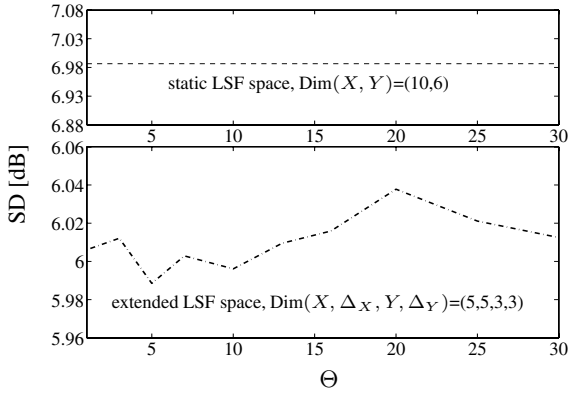

Fig. 3: *BWE performance with memory inclusion (bottom) versus performance with no memory inclusion (top).*

narrowband parameters.

Thus, we employ two GMMs to statistically model the correlation between the narrowband LSFs (and frame log-energy) and a) highband LSFs, and b) the excitation gain. Ideally, any benefit of incorporating memory into BWE should entail as little added complexity as possible. Accordingly, we evaluate BWE performance for an extended feature space that is obtained by replacing–rather than appending–half the static features by their delta coefficients, thus preserving dimensionality. Memory inclusion, thus implemented, consequently involves no added computational requirements or increase in amount of GMM training data. Noteworthy is that Table 1 further shows that replacing half the static features, rather than appending them, results in higher relative increase in highband certainty.

Hence, in conformity with the dimensionality of the feature spaces of Fig. 1, we use a static space of $\mathrm{Dim}(X,Y)$=(10,6) as the reference for BWE performance. For this static case, a GMM models the distribution of the wideband random vector (RV) consisting of 9 narrowband LSFs, the narrowband frame log-energy, and 6 highband LSFs, while the second GMM models that of the RV consisting of the 10 narrowband features as above, in addition to the excitation gain, g. The extended space case with $\mathrm{Dim}(X,\Delta_X,Y,\Delta_Y)$=(5,5,3,3) employs a GMM modelling the distribution of a wideband RV consisting of 4 narrowband LSFs, their 4 delta coefficients, the narrowband frame log-energy and its delta coefficient, 3 highband LSFs and their 3 delta coefficients. The second GMM models the same narrowband features in addition to the excitation gain, g.

We evaluate BWE performance in the missing 3.4–7 kHz band by the SD of the extended highband speech, given by:

$$\mathrm{SD}^2 = \frac{1}{\pi}\int_{\omega_l}^{\omega_h}\left(20\log_{10}\frac{g}{|A_Y(e^{j\omega})|}\cdot\frac{|\hat{A}_Y(e^{j\omega})|}{\hat{g}}\right)^2 \mathrm{d}\omega \quad (6)$$

where ω_l and ω_h are the cutoff frequencies of the missing high band, g and $A_Y(e^{j\omega})$ are the highband excitation gain and frequency spectrum of the original wideband signal, respectively, while $\hat{g}$ and $\hat{A}_Y(e^{j\omega})$ are the reconstructed highband excitation gain and frequency spectrum as estimated by GMMs.

Fig. 3 shows the improvement in BWE performance resulting from incorporating memory inclusion through delta features. The mere exploitation of delta features–regardless of how much memory is actually used to estimate them–clearly results in a considerable objective quality improvement of about 1 dB on average corresponding to 14% relative SD decrease, approximately. Inspecting Fig. 3 more closely for the effect of the extent of incorporated memory (represented by Θ), reveals that quality improvement is greatest for, roughly, $5 \lesssim \Theta \lesssim 10$ corresponding to $100 \lesssim t \lesssim 200$ ms, reaching a maximum SD relative decrease of 14.28%. This time range corresponds to triphone durations, asserting our conclusion in [3] that BWE schemes are expected to benefit mostly through such short-term memory inclusion. Expectedly, this range is further in accordance with that representing the highest gains in highband certainty per Fig. 1, further confirming the strong correlation between highband certainty and BWE reconstructed speech quality.

6. Conclusions and future work

We review the information theoretic justification of incorporating memory inclusion in BWE of narrowband speech. By using VQ, we improve our previous estimates in [3] of the effect of the amount of memory included in speech parametrization on the extended speech quality. We extend the analysis to LSFs, widely used in BWE schemes. We further translate the highband certainty gains obtained by memory inclusion into practical BWE performance improvement through embedding dynamic speech features in an LP-based BWE system. Objective analysis of reconstructed highband speech confirms the positive effect of memory inclusion on improving BWE performance. Our results in Section 4 lead us to further conclude that MFCCs are superior to LSFs in terms of retaining information mutual to both narrow and high speech frequency bands, thus making them the preferred means of parametrization in terms of the potential for improving BWE performance. Hence, the implementation of MFCC-based BWE will be the focus of our future work.

7. References

[1] M. Nilsson, S. V. Andersen and W. B. Kleijn, "On the mutual information between frequency bands in speech", in *Proc. ICASSP*, pp. 4085–4088, 2000.

[2] M. Nilsson, H. Gustafsson, S. V. Andersen and W. B. Kleijn, "Gaussian mixture model based mutual information estimation between frequency bands in speech", in *Proc. ICASSP*, pp. 525–528, 2002.

[3] A. H. Nour-Eldin, T. Z. Shabestary and P. Kabal, "The effect of memory inclusion on mutual information between speech frequency bands", in *Proc. ICASSP*, pp. III-53–56, 2006.

[4] P. Jax and P. Vary, "An upper bound on the quality of artificial bandwidth extension of narrowband speech signals", in *Proc. ICASSP*, pp. 237–240, 2002.

[5] Y. Qian and P. Kabal, "Combining equalization and estimation for bandwidth extension of narrowband speech", in *Proc. ICASSP*, pp. I-713–716, 2004.

[6] T. D. Lookabaugh and R. M. Gray, "High-resolution quantization theory and the vector quantizer advantage", *IEEE Trans. Inform. Theory*, vol. 35, pp. 1020–1033, 1989.

[7] S. P. Lloyd, "Least squares quantization in PCM", *IEEE Trans. Inform. Theory*, vol. 28, pp. 129–137, 1982.

[8] W. Fisher, et al., "The DARPA speech recognition research database: specification and status", in *Proc. DARPA Workshop on Speech Recognition*, pp. 93–99, 1986.

[9] R. Hagen, "Spectral quantization of cepstral coefficients", in *Proc. ICASSP*, pp. 509–512, 1994.

[10] P. Jax and P. Vary, "Feature selection for improved bandwidth extension of speech signals", in *Proc. ICASSP*, pp. 697–700, 2004.