AVOIDING OVER-ESTIMATION IN BANDWIDTH EXTENSION OF TELEPHONY SPEECH

Mattias Nilsson and W. Bastiaan Kleijn

Department of Speech, Music and Hearing
KTH (Royal Institute of Technology)
100 44 Stockholm, Sweden

ABSTRACT

In this paper we present a new way of treating the problem of extending a narrow-band signal to a wide-band signal. For many cases of bandwidth extension, the high-band energy is over-estimated, leading to undesirable audible artifacts. To overcome these problems we introduce an asymmetric cost-function in the estimation process of the high-band that penalizes over-estimates more than under-estimates of the energy in the high-band. We show that the resulting attenuation of the estimated high-band energy depends on the broadness of the a-posteriori distribution of the energy given the extracted information about the narrow-band. Thus, the uncertainty about how to extend the signal at the high-band influences the level of extension. Results from listening test show that the proposed algorithm produces less artifacts.

1. INTRODUCTION

A speech signal that has passed through the public switched telephony network (PSTN) has a characteristic low-pass quality. This is due to the limited bandwidth of the telephony channel (0.3 - 3.4 kHz). Transmitting wide-band speech (speech sampled at 16 kHz) would preserve the naturalness of the speech signal that is otherwise lost. However, changing the existing PSTN infrastructure to handle wide-band speech transmission would entail an enormous cost. Therefore, recovery of the wide-band speech signal by estimation of the high- (3.4 - 8 kHz) and low- (0.1 - 0.3 kHz) frequency bands from the telephony signal would be of great use. In [1] we showed that there is mutual information between the narrow- and high-band and that, thereby, it is justified to try to recover the high-band given the knowledge of the narrow-band.

The recovery of wide-band speech (0.1-8 kHz) given the narrow-band (0.3-3.4 kHz) speech is a challenging task and there have been a number of contributions on this topic. In [2] a simple multi-rate approach to the problem was presented and in, for instance, [3, 4] methods based on vector quantization for the mapping between narrow-band and wide-band were used. Statistical approaches have also been applied to the problem. The pioneering paper utilizing a statistical framework was the paper by Cheng et al. [5], a Gaussian Mixture Model (GMM) based method was used in [6], and a hidden Markov model (HMM) in [7].

Regardless of the bandwidth extension method used, a common problem is the introduction of artifacts in the extension bands that make the bandwidth extended signal often more annoying than the original narrow-band speech signal. These artifacts are generally due to over-estimation of the high-band energy.

Therefore, if a bandwidth extension algorithm is to be used in a real telephony system, it is important that no audible energy over-estimates of the signal in the high- and low- frequency bands are made. It is in that sense better to under-estimate the energy in such regions. This motivates the idea of having an asymmetric cost-function that penalizes over-estimates more than under-estimates of the energy in the high-band.

Our aim is to arrive at a confidence controlled bandwidth extension algorithm. That is, we want to be able to control the level of extension (energy, shape) depending on how confident we are about the estimates of the high-band parameters, such that the bandwidth extended signal is wide-band and contains no artifacts. This paper is a first step in this direction.

2. SYSTEM STRUCTURE

The core of our bandwidth extension is a GMM. The GMM models the joint probability density function, $f_Z(z)$, of the random variable feature vector, Z, and is of the form

$$f_Z(z) = \sum_{m=1}^{M} \alpha_m f_Z(z|\theta_m), \tag{1}$$

where M is the number of mixtures components, and α_m is the weight of mixture number m. The distribution $f_Z(z|\theta_m)$ is a multivariate Gaussian distribution,

$$f_Z(z|\theta_m) = \frac{1}{(2\pi)^{\frac{d}{2}}|C_m|^{\frac{1}{2}}} \exp\left(-(z-\mu_{z_m})^T C_m^{-1}(z-\mu_{z_m})\right), \tag{2}$$

with mean vector μ_m and covariance matrix C_m both collected in $\theta_m = \{\mu_m, C_m\}$, and d the feature vector dimension. Herein, the feature vector, z, has dimension 22 and consists of:

- a narrow-band spectral envelope part modeled by 15 linear frequency cepstral coefficients (LFCCs), $x = \{x_1 \dots x_{15}\}$,
- a high-band spectral envelope part modeled by 5 LFCCs, $y = \{y_1 \dots y_5\}$,
- an energy-ratio variable g being the difference in log energy between the high- and narrow-band, i.e., $g = y_0 - x_0$,
- a measure on the degree of voicing, r.

The degree of voicing is determined by the maximum of the normalized autocorrelation function within the lag range corresponding to 50-400 Hz, i.e.,

$$r = \max_{20 \le \tau \le 160} \frac{\sum_{n=0}^{N-1} s(n)s(n+\tau)}{\sqrt{\sum_{k=0}^{N-1} s(k)^2 \sum_{l=0}^{N-1} s(l+\tau)^2}}, \tag{3}$$

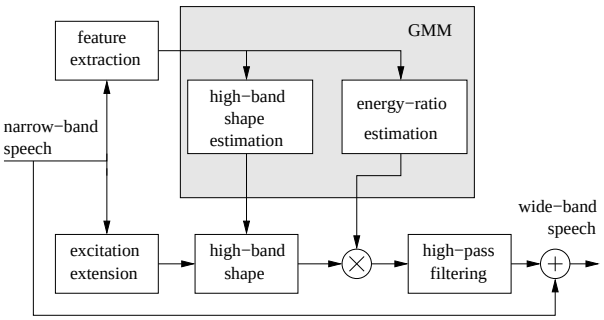

Fig. 1. System architecture.

where $s = s(1) \dots s(160)$ is a 20 ms narrow-band speech segment sampled at 8 kHz. The model parameters α_m and θ_m, for $m = 1 \dots M$, were obtained using the EM-algorithm described in [8] on a training set extracted from the TIMIT database [9]. The size of the training set was selected to be 100,000 non-overlapping 20 ms segments from which the features were calculated. In our experiments, we used 32 mixture components ($M = 32$).

In contrast to the GMM approach presented in [6], we have chosen to use diagonal covariance matrices in our model. We have done so mainly for three reasons:

1. A GMM with full covariances can be modeled by a higher-order GMM with diagonal covariances [10].
2. Under the Gaussian assumption, the covariance matrix of the cepstral coefficients is approximately diagonal [11].
3. Significantly fewer parameters have to be estimated.

A schematic of our bandwidth extension scheme is depicted in Figure 1. From the narrow-band speech, 15 cepstral coefficients, x, and the degree of voicing, r, are derived. Then, using an asymmetric cost-function together with the a-posteriori distribution of the energy-ratio given the narrow-band shape and narrow-band voicing parameter, we obtain an estimate of the energy-ratio between the narrow- and high-band. The asymmetric cost-function penalizes over-estimates of the energy-ratio more than under-estimates. Furthermore, as is shown in section 2.1, a narrow a-posteriori distribution of the energy-ratio results in less penalty on the energy-ratio than a broad distribution. The energy-ratio estimate, together with the narrow-band shape and the degree of voicing, form a new a-posteriori distribution of the high-band shape and an MMSE estimate of the high-band envelope is computed. A modified spectral folding excitation is then filtered with the energy-ratio controlled high-band envelope and added to the narrow-band to form a wide-band speech signal.

2.1. Energy-ratio estimation

The most common cost-function used for parameter estimation from a conditioned probability function is the quadratic. If we use the a-posteriori distribution of the energy-ratio given the narrow-band shape and degree of voicing together with a quadratic cost-function, we obtain the standard MMSE estimate, i.e.,

$$\begin{aligned} \hat{g}_{MMSE} &= \arg\min_{\hat{g}} \int_{\Omega_g} (\hat{g} - g)^2 f_{G|XR}(g|x, r) dg \\ &= E[G|X = x, R = r] \end{aligned}$$

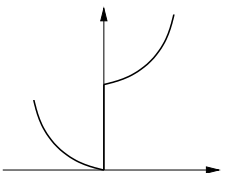

Fig. 2. The asymmetric cost-function.

$$\begin{aligned} &= \int_{\Omega_g} g \frac{\sum_{m=1}^{M} \alpha_m f_{GXR}(g, x, r|\theta_m)}{\sum_{k=1}^{M} \alpha_k f_{XR}(x, r|\theta_k)} dg \\ &= \sum_{m=1}^{M} w_m(x, r) \int_{\Omega_g} g f_{G|XR}(g|x, r, \theta_m) dg \\ &= \sum_{m=1}^{M} w_m(x, r) \int_{\Omega_g} g f_G(g|\theta_m) dg \\ &= \sum_{m=1}^{M} w_m(x, r) \mu_{g_m}, \end{aligned} \quad (4)$$

where in the second last step, we used the fact that each individual mixture component has a diagonal covariance matrix and, thus, independent components. Since we perceive over-estimates of the energy ratio as more annoying than under-estimates, we should use an asymmetric cost-function instead of a symmetric one in order to penalize over-estimates more than under-estimates. The asymmetric cost-function we use is shown in Figure 2 and is expressed as,

$$C = bU(\hat{g} - g) + (\hat{g} - g)^2, \quad (5)$$

where $bU(\cdot)$ is the step function with amplitude b. The amplitude of the step can be seen as a tuning parameter, giving a possibility to control the degree of extension. The estimate of the energy-ratio can then be written as,

$$\hat{g} = \arg\min_{\hat{g}} \int_{\Omega_g} (bU(\hat{g} - g) + (\hat{g} - g)^2) f_{G|XR}(g|x, r) dg. \quad (6)$$

The estimate is found by differentiating the right-hand side of (6) and set it equal to zero. Assuming that the order of differentiation and integration may be interchanged we can write the derivative of (6) as,

$$\begin{aligned} \sum_{m=1}^{M} w_m(x, r) \int_{\Omega_g} (b\delta(\hat{g} - g) + 2(\hat{g} - g)) f_G(g|\theta_m) dg &= 0, \\ \sum_{m=1}^{M} w_m(x, r) b f_G(\hat{g}|\theta_m) + 2\hat{g} - 2 \sum_{m=1}^{M} w_m(x, r) \mu_{g_m} &= 0, \end{aligned}$$

which finally yields,

$$\hat{g} = \sum_{m=1}^{M} w_m(x, r) \mu_{g_m} - \frac{b}{2} \sum_{m=1}^{M} w_m(x, r) f_G(\hat{g}|\theta_m), \quad (7)$$

which has to be solved using a numerical technique, e.g., by means of a grid search. As shown in (7) the energy-ratio estimation depends on the shape of the posterior distribution. Thus, the penalty

on the minimum mean squared error estimate of the energy ratio depends on the width of the posterior distribution. If the a-posteriori distribution $f_{G|XR}(g|x,r)$ is narrow, we are more confident on the MMSE estimate than if the a-posteriori distribution is broad.

2.2. High-band envelope estimation

In order to have a probable combination of the high-band shape and energy-ratio, the high-band shape estimation is performed by conditioning on the energy-ratio estimate, the narrow-band shape, and the degree of voicing in the narrow-band speech segment. By using a GMM with diagonal covariance matrices, the MMSE estimate of the high-band shape simply becomes,

$$\begin{aligned}\hat{y}_{MMSE} &= E[Y|X=x, R=r, G=\hat{g}] \\ &= \sum_{m=1}^{M} \frac{\alpha_m f_{XRG}(x,r,\hat{g}|\theta_m)\mu_{y_m}}{\sum_{n=1}^{N}\alpha_n f_{XRG}(x,r,\hat{g}|\theta_n)}. \end{aligned} \quad (8)$$

Thus, the estimate of the high-band is a weighted sum of mean vectors from the different mixtures, where the weighting is simply the probability of a certain mixture component given the narrow-band feature vector and the estimate of the energy-ratio. This is closely related to the method described in [5].

2.3. Excitation extension

For the extension of the narrow-band excitation, we use a modification of spectral folding technique in [12]. Instead of folding the spectrum around 4 kHz, which due to the telephony channel would result in very low energy in the region between 3.4 and 4.6 kHz, we cut out the part between 2 and 3 kHz of the spectrum of the narrow-band excitation and repeatedly fold this segment (starting at 3 kHz) until we have covered the frequency region of interest. To avoid an overly periodic excitation at the higher frequencies for voiced segments, we let the constructed folded spectrum smoothly evolve in frequency to a white noise spectrum. The amplitude of white noise was chosen to be equal to the mean of the amplitude spectrum of the narrow-band excitation. The transition between the periodic and noise region was set ad-hoc such that above 6 kHz the noise spectrum dominates totally.

3. SIMULATIONS

Our simulations are intended to show two important properties. The first one is that the asymmetric cost-function penalizes the energy-ratio estimates more when the a-posteriori distributions of the energy-ratio is wide, than when the a-posteriori distribution is narrow. The second property is that regions with wide conditional energy-ratio distributions are potential regions where the standard MMSE technique over-estimates the energy-ratio. In the simulations, we bandwidth extend narrow-band speech files, using both the standard MMSE for the estimation of the high-band shape and energy-ratio and using the proposed method with the asymmetric cost-function. None of the test speech files were included in the training set.

3.1. Asymmetric cost-function

To evaluate the proposed scheme, we analyzed the behavior of the asymmetric cost-function depending on the a-posteriori distribution of the energy-ratio given the narrow-band shape and degree of voicing. For wide a-posteriori distributions, the asymmetric cost-function imposes a larger penalty than for narrow a-posteriori distribution. This is shown in Figure 3, where the wide a-posteriori distribution results in approximately 11 dB attenuation of the MMSE estimated energy-ratio compared to approximately 6 dB attenuation for the narrow distribution (the attenuation is the distance between the minima of the asymmetric and squared error cost in the right sub-plots in Figure 3). Note that, the narrower the a-posteriori distribution is, the closer is the minimum of the asymmetric cost-function to the minimum of the squared error cost-function.

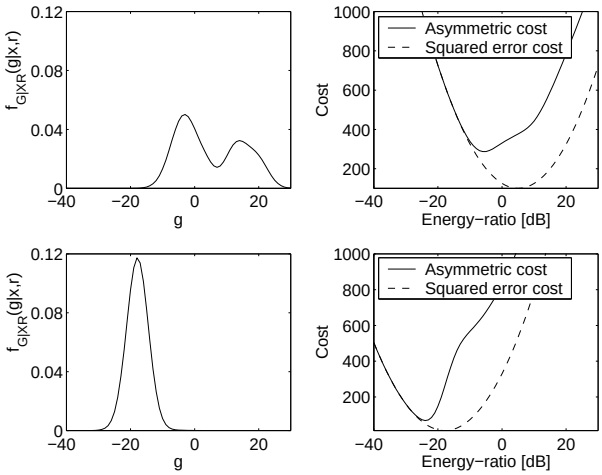

Fig. 3. The upper left and lower left plots show two examples of the distribution $f_{G|XR}(g|x,r)$ and the upper right and lower right plots show the associated cost. Note that, in the upper right plot the distance between the minima of the asymmetric- and the squared error cost is greater than the distance between the minima in the lower right plot.

3.2. Energy-ratio over-estimation

The bandwidth extended spectra corresponding to the wide and narrow distributions, shown in Figure 3, are shown in Figure 4 and 5, respectively. From Figure 4, which shows an unvoiced segment, it is clear that the MMSE method produces an over-estimate of the energy-ratio between the narrow- and high-band, that will result in an audible artifact. The voiced segment shown in Figure 5 shows, that even though the a-posteriori distribution was narrow, we estimated an energy-ratio that is lower than the energy-ratio for the original signal. However, for voiced segments the under-estimates of the energy-ratio are not that noticeable. For unvoiced regions with MMSE estimates of the energy-ratio that are below the true energy ratio, the proposed algorithm lowers the energy ratio even further, thus creating a signal with less of a wide-band character. Nevertheless, the bandwidth extended signal produced by the proposed algorithm still has a clear wide-band sensation when compared to the narrow-band signal.

4. LISTENING TEST

To evaluate the proposed algorithm we conducted a listening test where the subject was asked to grade the degree of artifacts in the

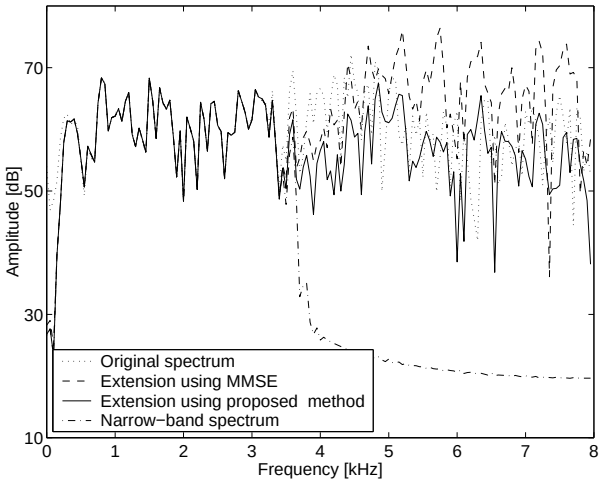

Fig. 4. Bandwidth extended segment where the energy-ratio is estimated from a wide a-posteriori distribution. Hamming windowed 20 ms segments were used in the computation of the spectra.

	Degree of artifacts			
Method	1.Severe	2.Moderate	3.Minor	4.None
MMSE	4.9 %	40.6 %	44.7 %	9.8 %
Proposed	0.4 %	17.9 %	49.6 %	32.1 %

Table 1. Results from the listening test.

bandwidth extended signal. The same test was then performed using the standard MMSE estimation of the high-band parameters. The listener could choose from four ratings with respect to the artifacts: 1. Severe, 2. Moderate, 3. Minor, and 4. None. In the listening test, the proposed algorithm was used with the step of the asymmetric cost-function set to 500 ($b = 500$). Seven subjects were used in the listening test and the results are shown in Table 1. From the listening test, we conclude that the level of artifacts have been reduced by a significant amount.

5. CONCLUDING REMARKS

The proposed algorithm for bandwidth extension showed by the listening test to have less artifacts. The asymmetric cost-function makes it possible to control the level of extension. An implicit effect of the asymmetric cost-function is that it penalizes regions where the estimate of the energy-ratio is less certain (wide a-posteriori distribution) more than regions with more certain energy-ratio estimate (narrow a-posteriori distribution). The existing scheme can be considered as conservative in its bandwidth extension. However, the proposed scheme produces an extended signal which is clearly wide-band compared to the narrow-band speech signal.

6. REFERENCES

[1] M. Nilsson, S. V. Andersen, and W. B. Kleijn, "On the mutual information between frequency bands in speech," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, 2000, pp. 1327–1330.

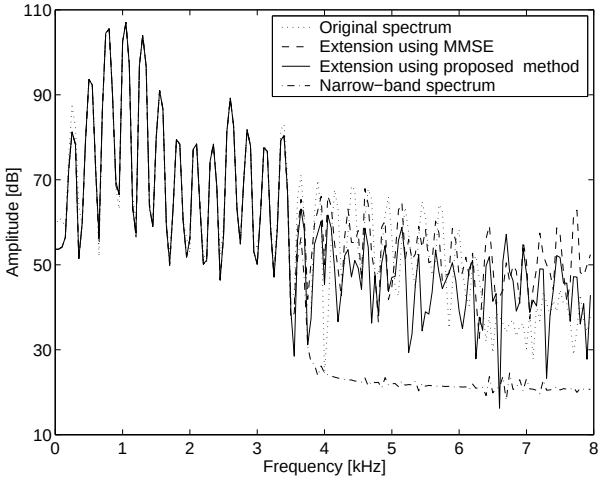

Fig. 5. Bandwidth extended segment where the energy-ratio is estimated from a narrow a-posteriori distribution. Hamming windowed 20 ms segments were used in the computation of the spectra.

[2] H. Yasukawa, "Quality enhancement of band limited speech by filtering and multirate techniques," in *Proc. Int Conf. Speech Lang. Process.*, 1994, pp. 1607–1610.

[3] N. Enbom and W. B. Kleijn, "Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients," in *IEEE Workshop on Speech Coding*, Porvoo,Finland, 1999, pp. 1953–1956.

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

[5] Y. M. Cheng, D. O'Shaugnessy, and P. Mermelstein, "Statistical recovery of wideband speech from narrowband speech," *IEEE Trans. Speech and Audio Proc.*, vol. 2, no. 4, pp. 544–548, 1994.

[6] K. Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, 2000, pp. 1843–1846.

[7] P. Jax and P. Vary, "Wideband extension of telephone speech using a hidden markov model," in *IEEE Workshop on Speech Coding*, Delavan,USA, 2000, pp. 130–135.

[8] A. P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Soc., Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.

[9] DARPA-TIMIT, "Acoustic-phonetic continuous speech corpus," *NIST Speech Disc 1-1.1*, 1990.

[10] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech and Audio Proc.*, vol. 3, no. 1, pp. 72–83, 1995.

[11] Y. Ephraim and M. Rahim, "On second order statistics and linear estimation of cepstral coefficients," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, 1998, pp. 965–968.

[12] J. Makhoul and M. Berouti, "High-frequency regeneration in speech coding systems," in *Proc. IEEE Int. Conf. Acoust. Speech Sign. Process.*, 1979, pp. 428–431.