Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (Montreal, QC), pp. I-713-I-716, May 2004

COMBINING EQUALIZATION AND ESTIMATION FOR BANDWIDTH EXTENSION OF NARROWBAND SPEECH

Yasheng Qian *Peter Kabal*

Department of Electrical and Computer Engineering
McGill University, Montreal, Canada

ABSTRACT

Current public telephone networks compromise voice quality by bandlimiting the speech signal. Telephone speech is characterized by a bandpass response from 300 to 3400 Hz. The voice quality is perceived as being much worse than for wideband speech (50–7000 Hz). We present a novel approach which combines equalization and estimation to create a wideband signal, with reconstructed components in the 3400 Hz to 7000 Hz range. Equalization is used in the 3400-4000 Hz range. Its performance is better than statistical estimation procedures, because the mutual dependencies between the narrowband and highband parameters are not sufficiently large. Subjective evaluation using an Improvement Category Rating shows that the reconstructed wideband speech using both equalization and estimation substantially enhances the quality of telephone speech. We have also evaluated the performance on the narrowband output of several standard codecs. Overall, the use of equalization for part of the highband regeneration makes the system more robust to phonetic variability and speaker gender.

1 Introduction

Voice quality is compromised by bandlimiting speech signals. In current public telephone networks, the effective upper band boundary of 3400 Hz gives high sentence intelligibility (up to 99%). The intelligibility of individual syllables is about 90%, but is much lower for unvoiced phonemes such as /s/ and /f/, because their spectra go well beyond 3400 Hz. The low frequency cutoff in telephony, 300 Hz, is set to suppress the power line longitudinal interference and other low frequency electrical noises. Typically, there is more than 25 dB attenuation at 50–60 Hz. As a result of the bandlimiting, telephone speech sounds very different from broadcast speech (the *de facto* definition of wideband speech, with a bandwidth of 50–7000 Hz). The loss in bandwidth compromises naturalness, fidelity and intelligibility.

A number of researchers [1–5] have used estimation methods, based on a speech production model, to restore the missing frequency components. The probabilistic estimation of the highband spectrum envelope and energy relies on the mutual statistical dependencies between the available narrowband spectrum and the missing spectrum. The larger the dependencies, the better the estimates. The mutual dependencies between the narrowband region and the high frequency region have been investigated from an information theory perspective using mutual information and the differential entropy measures [6]. The small mutual information (about 1.5 bits) implies that inevitably the reconstructed high frequencies have a large spectral error.

We have reviewed published objective spectral error figures, along with our own RMS log spectral distortion (RMS-Log-SD) measurements. These have been obtained with different estimation algorithms, including VQ mapping, GMM mapping and HMM mapping. The average spectral error in the range 4000–8000 Hz seems to be about 6.0 dB and changes 1–2 dB with different parameter options, such as VQ codebook size, the number in GMM components and the dimension of the state in HMM approaches. The spectral error is much higher than the threshold of 1 dB, which is usually considered as a threshold for spectral transparency in the narrowband speech. However, the frequency components above 4000 Hz play a much smaller perceptual role than those in the lowband. We have found that even with a mean RMS-Log-SD of 6 dB, we can achieve high quality reconstructed wideband speech.

The quality of the excitation signal used for resynthesizing the highband is another important factor which affects the quality of the restored speech. The excitation can be modelled in a number of ways, including spectral folding, non-linear operations and harmonic shifting.

Our new approach employs equalization, as much as possible, 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 4000 Hz 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 can be used to produce an enhanced excitation signal which will be used in the region above 4000 Hz. Statistical estimation is used to generate the complementary spectrum in the range from 4000 to 7000 Hz. The use of equalization has made a major contribution to voice quality improvement in our bandwidth extension scheme.

2 Equalization

We characterize the speech signal after passing an ITU-T G.712 channel filter as follows: (1) The channel filter attenuates the speech signal from 0 dB to 18 dB between 3400 Hz and 4000 Hz and from 0 dB to 10 dB in the frequency between 300 Hz to 100 Hz. We refer to those frequency

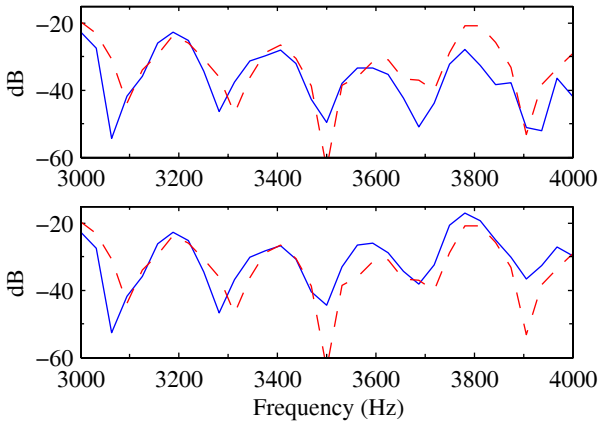

Fig. 1 Comparisons of spectra (3000–4000 Hz) for a voiced frame, (top) without equalization and (bottom) with equalization. The dashed line represents the wideband speech spectrum.

components as attenuated components. Those attenuated components can be restored by equalization. (2) Components above 4000 Hz are missing due to sampling at 8 kHz. We refer to these as lost components. These lost components can be reconstructed only by statistical estimation. (3) Below 100 Hz, there is a deep valley at 50–60 Hz with more than 25 dB attenuation. We do not attempt to restore those components.

We have designed two equalizers to recover the attenuated components. The first equalizer has a boost of 10 dB from 3800 Hz to 4000 Hz. The second one gives a gain of 10 dB at 100 Hz. The frequency response of the equalized channel filter is almost flat from 100 Hz to 3850 Hz. Figure 1 shows that the first equalizer has restored the attenuated spectrum of a voiced frame in the frequency range of 3400 to 4000 Hz. We have observed that the voice quality of the equalized speech is noticeably better than narrowband speech. Although the equalized speech still resides in the frequency range from 100 Hz to 4000 Hz, it plays important role for reconstructing the lost components over 4000 Hz. The enhanced excitation, the spectrum envelope and the excitation gain estimation for the lost band will be generated from the equalized speech.

3 Excitation Generation

In our previous work [3], we have used bandpass modulated Gaussian noise (BP-MGN) derived from a bandpass region of the narrowband speech as the highband excitation. The BP-MGN has proved to be an excellent excitation in many cases. However, the BP-MGN approach does not work well for some phonemes that have weak responses in the region 2–3 kHz. This is illustrated in Fig. 2. The resulting BP-MGN excitation does not contain sufficient highband components. That results in severely distorted reconstructed highband components as shown in the middle trace of the Fig. 3. We have replaced the 2–3 kHz bandpass filter by a 3–4 kHz bandpass filter in the excitation generation. The equalized speech is used in the EBP-MGN generation. Because there are now richer components in the 3–4 kHz band,

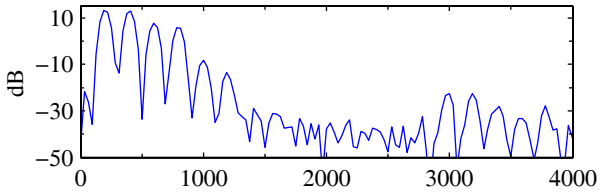

Fig. 2 The spectrum of a voiced phoneme with weak components between 2 kHz and 3 kHz.

the excitation can produce an adequate highband components for most phonemes. The EBP-MGN approach makes the excitation more robust to differences in phonemes and speaker gender. We pass upsampled narrowband speech through a 3–4 kHz bandpass filter. The bandpass signal is

$$s_{bp}(n) = s_{bb}(n)\cos(2\pi f_o n). \quad (1)$$

where $f_o = 3.5$ kHz and $s_{bb}(n)$ is a baseband signal. The envelope of the bandpass signal is $|s_{bp}(n)|$. The spectrum of the envelope is $S_{bpe}(\omega)$, which contains pitch harmonics in the highband due to non-linear operation on the bandpass signal. The EBP-MGN excitation, $e(n)$ is a bandpass-envelope modulated by a Gaussian noise. Figure 3 (top) shows the original spectrum of the highband components of a voiced phoneme. Figure 3 (middle) shows the spectrum of the reconstructed highband components using BP-MGN excitation. Figure 3 (bottom) shows the recovered highband components with EBP-MGN excitation. The reconstructed signal components with EBP-MGN excitation shows better reproduction in the range between 4800 Hz and 7000 Hz. The excellent regeneration occurs in the frequency region from 3400 Hz to 4000 Hz where equalization is applied.

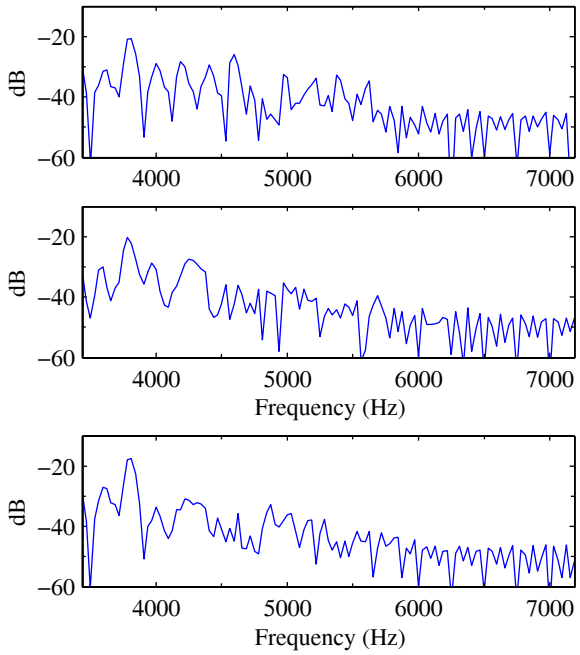

Fig. 3 The highband spectra of a voiced phoneme showing differences between excitation generation methods. The original spectrum (top), the reconstructed highband spectrum with BP-MGN (middle); with EBP-MGN (bottom).

Our listening tests confirm that the EBP-MGN excitation works better than BP-MGN as a substitute for the highband excitation.

4 Estimation of the Excitation Gain

An excitation gain, g, is introduced to scale the synthesized highband components to an appropriate energy. The energy of the reconstructed highband components should ideally be equal to the energy of the corresponding frequency band in wideband speech. The reconstructed highband signal, $s_{res}(n)$, is the convolution of the highband excitation $e(n)$, multiplied by g and the impulse response of the LP synthesis filter, $h(n)$, that is,

$$s_{res}(n) = g\,[e(n) * h(n)]. \tag{2}$$

The excitation gain g is calculated as the square root of the energy ratio of the original highband signal, $s_{hp}(n)$, to the resynthesized one, $s_{res}(n)$, of each frame.

$$g = \sqrt{\frac{||s_{hp}(n)||^2}{||s_{res}(n)||^2}}. \tag{3}$$

The true value of the excitation gain can only be determined during training.

In a training system, the wideband speech first passes through a lowpass and a highpass filter. Both narrowband and highband components, s_{lp} and s_{hp} are then input to an LPC analysis stage to get narrowband and highband spectrum parameters. With the excitation EBP-MGN and the highband spectrum parameters, the highband components can be synthesized.

The excitation gain g is a random variable, which can not directly be determined from narrowband speech. However, we assume that g is, to certain degree, correlated with narrowband spectrum and pitch gain, so that it can be statistically estimated from narrowband parameters. We first derive the statistical parameters of a Gaussian mixture pdf, which is a joint pdf of the three parameters, the narrowband spectrum, pitch gain and the excitation gain from the training program. Then, we use probabilistic estimation to get a g estimate on Minimum Mean Square Error criterion.

Because of the well-known properties (ordering and quantization error resilience) of the Line-Spectrum-Frequencies (LSF) representation, we employ 14 and 10 LSFs to represent the narrowband and highband spectrum, respectively. The LSFs, β and the excitation gain, g, are part of a random vector, whose probability density function (pdf) can be approximated by a GM pdf.

The GM pdf is a weighted sum of M D-dimensional joint Gaussian density distributions.

$$p_Z(\boldsymbol{z}|\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{i=1}^{M} \alpha_i b_i(\boldsymbol{z}|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i). \tag{4}$$

where M is the number of individual Gaussian components, the $\alpha_i,\ i = 1,\ldots,M$ are the (positive) mixture weights, and Z is a D-dimensional random vector. Each density is a D-variate Gaussian pdf of the form,

$$b_i(\boldsymbol{z}|\mu_i,\boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp(\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_i)). \tag{5}$$

with mean vector $\boldsymbol{\mu}_i$, and covariance matrix $\boldsymbol{\Sigma}_i$. The GM pdf is defined by the mean vectors, the covariance matrices and the mixture weights for the Gaussian components.

The parameter set, $\{\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma}\}$ can be estimated by the maximum likelihood (ML) method. The ML algorithm finds the GM pdf parameters with maximum probability density for the given training data. We employ the popular expectation-maximization (EM) algorithm [7] to determine the set of GM density parameters iteratively.

The training data of wideband speech are taken from Speech Database with a total of 150 000 frames each of 20 ms with 1 320 utterances, spoken by 24 speakers (half male and half female).

Z is a 16-dimensional random vector, representing the 14 narrowband LSFs, the excitation gain g and the pitch gain, β. The number of mixtures, M, is 128. The covariance matrices, $\boldsymbol{\Sigma}_i$, are diagonal. The $\hat{g}$ estimate is based on the GM joint density distribution of Eq. (4). Let the random vector $\boldsymbol{x}$ be the combination vector of the narrowband LSF vector and the pitch gain β. For a given estimate, $\hat{g}$, the mean-square error is

$$\varepsilon^2 = \int_g ||g-\hat{g}||^2 p_{g|X}(g|\boldsymbol{x})\,dg. \tag{6}$$

The optimal estimate which minimizes the error is found from $\partial\varepsilon^2/\partial\hat{g} = 0$.

$$\hat{g}_{opt} = \frac{\int_g g\,p_{g|X}(g|\boldsymbol{x})\,dg}{\int_g p_{g|X}(g|\boldsymbol{x})\,dg} = \frac{\sum_{i=1}^{M}\alpha_i b_i(\boldsymbol{x})\mu_{ig}}{\sum_{j=1}^{M}\alpha_j b_j(\boldsymbol{x})}. \tag{7}$$

where μ_{ig} is the mean of g of the i-th Gaussian component. The $\hat{g}_{opt}$ estimate is the conditional expectation of the μ_{ig} mixture mean, given a narrowband LSF vector and a pitch gain. Similarly, we have established a GMM for the narrowband LSF vector, the pitch gain and the highband LSF vector. The highband LSF vector can be estimated with an equation similar to Eq. (7). We take pitch gain as an extra parameter, because an acoustic-phonetic classification based on the pitch gain, β, was beneficial in our previous work.

The gain-estimation-ratio, Rg_{est} quantifies the the ratio of the true gain and the estimated gain in dB.

$$Rg_{est} = 20\log_{10}(g/\hat{g}_{opt}). \tag{8}$$

Having Rg_{est} larger than zero means that the estimated gain value is lower than the true value in that frame. Figure 4 (top) shows the gain-estimation-ratio in (dB) for a female speaker. Figure 4 (bottom) represents the gain-estimation-ratio for a male speaker.

We observed that Rg_{est} is more likely to be larger than zero for female speech, while they are more likely to be less than zero for male speech. In both cases, the values fluctuate from negative to positive values. For example, typical female speech has 75% frames with Rg_{est} larger than zero. Male speech typically has 30% frames with Rg_{est} larger than zero. The gender discrepancy points out that the parameters we have used for the gain estimation GMM are not sufficient to model gender differences.

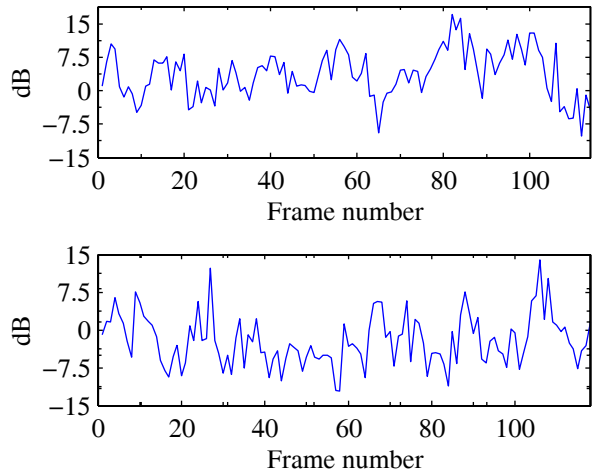

Fig. 4 The gain-estimation-ratio Rg_{est} in (dB); female speaker (top); male speaker (bottom).

5 Quality Evaluation

We have carried out several objective and subjective evaluations of the voice quality of our bandwidth extension system with equalization and estimation. We have measured the mean RMS-Log-SD in the missing highband (4–8 kHz). The definition of RMS-Log-SD is as follows:

$$\mathrm{SD}^2 = \frac{1}{\pi}\int_{\omega_l}^{\omega_h} 20\log_{10}\left(\frac{g}{\frac{|A_{\mathrm{hb}}(e^{j\omega})|}{g_{\mathrm{gmm}}}{|A_{\mathrm{gmm}}(e^{j\omega})|}}\right)^2 d\omega. \qquad (9)$$

where ω_l and ω_h are the cut-off frequencies of the lost band; g and g_{gmm} are the real excitation gain and the GMM-estimated excitation gain; $|A_{\mathrm{hb}}(e^{j\omega})|$ is the magnitude of the inverse filters of the highband signals of the wideband speech; $|A_{\mathrm{gmm}}(e^{j\omega})|$ is the estimated highband magnitude of response using the GMM parameters.

Table 1 RMS-LogSD in (dB)

	Female 1	Female 2	Male 1	Male 2
Mean	6.16	5.27	5.30	5.56
σ	3.42	2.87	2.60	3.31
Outliers 10 dB	17.5%	7.14%	5.93%	12.9%
Outliers 15 dB	0.87%	0.79%	0%	0.86%

Although the RMS-LogSD values in Table 1 are much larger than the threshold of transparency for narrowband spectra (1 dB), our listening tests show that RMS-LogSD of about 6.0 dB still can deliver high quality of reconstructed wideband speech.

We have used an Improvement Category Rating (ICR) to quantify the subjective quality of the bandwidth extended speech. A group of 20 people have participated the A/B comparison evaluation. A is the narrowband speech, while B is the bandwidth extended speech. The subjects have been asked to classify the difference between those two stimuli on an four-point quality scale, as listed in the Table 2.

We have also applied the A/B comparison tests using several standard coders. The codecs tested are ITU -T G.711 μ-law, G.723.1 MP-MLQ (6.3 kbits/s), G.729 CS-ACELP (8 kbits/s), ETSI AMR EFR (12.2 kbits/s) and IS-641 (7.4 kbits/s) codecs. Their ICR ratings are listed in Table 3. The listening evaluation shows that both uncoded and the G.711 coded telephone speech in present digital PTSNs show an ICR over 2.0. These systems gain a substantial benefit with bandwidth extension. The algorithm is also robust to the speaker gender. One early worry was that the equalization would bring up quantization noise for G.711 coding. We have ascertained that the equalization does not unduly emphasize quantization noise for G.711 μ-law coding.

Table 2 ICR for subjective evaluation

ICR	Condition
3	B is much better than A
2	B is better than A
1	B is slightly better than A
0	B is the same as or worse than A

Table 3 ICR for Different Codecs

Codec type	No codec	G.711	G.723.1
ICR	2.15	2.01	0
Codec type	G.729	AMR EFR	IS-641
ICR	1.5	1.7	1.2

References

[1] B. Iser and G. Schmidt, "Neural Networks Versus Codebooks in an Application for Bandwidth Extension of Speech Signals", *Proc. 8th European Conf. Speech, Commun. Tech.*, pp. 565–568, Sept. 2003.

[2] P. Jax and P. Vary, "On Artificial Bandwidth Extension of Telephone Speech", *Signal Processing*, vol. 83, pp. 1707–1719, Aug. 2003.

[3] Y. Qian and P. Kabal, "Dual-Mode Wideband Speech Recovery from Narrowband Speech", *Proc. 8th European Conf. Speech, Commun. Tech.,* pp. 1433–1437, Sept. 2003.

[4] M. Nilsson, H. Gustafsson, S. V. Anderson and W. B. Kleijn, "Gaussian Mixture Model based Mutual Information Estimation between Frequency Bans in Speech", *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, pp. 525–528, May 2002.

[5] K. P. Park and H. S. Kim, "Narrowband to Wideband Conversion of speech using GMM-based transformation", *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, pp. 1847–1850, June 2000.

[6] P. Jax and P. Vary, "An Upper Bound on the Quality of Artificial Bandwidth Extension of Narrowband Speech Signals", *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, pp. 237–240, May 2002.

[7] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *J. Royal Statistical Soc.*, Series B, vol. 39, pp. 1–38, 1977.