SPEECH ENHANCEMENT USING STC-BASED BANDWIDTH EXTENSION

J. Epps and W. H. Holmes

Department of Telecommunications, School of Electrical Engineering
The University of New South Wales 2052 Australia
E-mail: J.Epps@unsw.edu.au, H.Holmes@unsw.edu.au

ABSTRACT

Telephone speech is typically bandlimited to 4 kHz, resulting in a 'muffled' quality. Coding speech with bandwidth greater than 4 kHz reduces this distortion, but requires a higher bit rate to avoid other types of distortion. An alternative to coding wider bandwidth speech is to exploit correlation between the 0-4 kHz and 4-8 kHz speech bands to re-synthesize wideband speech from narrowband speech. This paper presents a method for re-synthesizing narrowband coded speech using sinusoidal transform coding (STC), modified codebook mapping and a novel method for the synthesis of highband unvoiced components. Informal listening test results indicate that this method produces a significant quality improvement in speech which has been coded using narrowband standards.

1. INTRODUCTION

Telephone speech (referred to hereafter as 'narrowband speech') has historically been bandlimited to 4 kHz, a bandwidth which represents a reasonable compromise between speech quality and transmission bandwidth for voiced speech, but often a poor one for unvoiced speech. There is an ongoing requirement for improved speech quality in mobile and fixed telephone services, and a particular need for wideband communications in such applications as talk-back radio and hands-free and internet telephony.

High quality wider bandwidth speech is best achieved by sampling speech at a higher rate, and allocating more bits for its transmission. This approach may not be desirable either due to bit rate constraints, or due to the huge base of narrowband PSTN already installed. An alternative to increasing the bit rate is to employ wideband enhancement of the narrowband speech, taking advantage of two types of correlation between the 0-4 kHz and 4-8 kHz bands of the speech signal:

- During voiced speech, the spectral fine structure is approximately harmonic
- The shape of the low band short-term spectral envelope determines to some extent the shape of the highband envelope

In this paper, some existing methods are considered and a new method is introduced. Section 2 reviews selected literature on wideband extension research to date, section 3 outlines the proposed wideband enhancement scheme, and results of its implementation are detailed in section 4.

2. PREVIOUS WIDEBAND ENHANCEMENT RESEARCH

2.1 Highband excitation generation

The first researchers to address the problem of highband spectral fine structure generation were Makhoul and Berouti [9], who proposed spectral folding and spectral translation methods for high frequency regeneration. Spectral folding in particular has been subsequently employed by many researchers [e.g. 3, 12, 16, 17], presumably for its simplicity, despite the lack of harmonicity in the resulting highband fine structure. Patrick and Xydeas [13] used a variety of non-linear transfer functions to produce high frequency fine structure. None of these methods guarantees spectral flatness (before highband spectral shaping).

Sinusoidal transform coding models the speech signal as a sum of sinusoids of arbitrary frequencies, amplitudes and phases, and can be used to synthesize high quality speech [11] up to any desired bandwidth. Spectral flatness is not an issue in STC synthesis since the sinusoid amplitudes are determined directly by the spectral envelope. Chan and Hui [4] re-synthesized fully decoded CELP (code excited linear prediction) speech using a multi-band excitation (MBE) [6] based approach, obtaining good results from subjective listening tests.

Generation of unvoiced components in wideband excitation has universally been achieved using bandpass random excitation. Little attention has been paid to the requirement for mixed (i.e. periodic and random) excitation in wideband enhancement. Many researchers have assumed that the highband is entirely unvoiced, and only one paper reviewed [4] considers the possibility of a mixed model for highband excitation.

2.2 Highband envelope generation

Approaches to the problem of highband spectral envelope generation have included adjustments of the highband gain [13, 14, 16], 2nd order LP coding of the 4-8 kHz band [15], time trajectory filtering of LP-cepstral coefficients [2], and statistical recovery of the highband envelope (with training based upon the EM algorithm) [5]. A promising approach is codebook mapping, which has achieved considerable gains

in speech quality enhancement [3, 4, 12, 17]. In this scheme a one-to-one mapping is applied to a narrowband envelope vector, transforming it to a wideband envelope vector which is subsequently used to synthesize the spectral envelope for the highband.

This paper describes a new wideband enhancement approach which combines the sinusoidal model of [11] with a novel model for highband mixed excitation and an improved highband envelope model.

3. STC-BASED WIDEBAND ENHANCEMENT

3.1 Overview

The overall scheme of the wideband re-synthesis method presented in this paper is illustrated in the block diagram of Fig. 1. The 0-3.8 kHz portion of the input narrowband speech was not re-synthesized since this spectral region was assumed to contain communication quality speech already.

3.2 Narrowband analysis

Analysis of the narrowband speech was performed using a Hamming window of 20 ms duration. For each frame of input speech, the 10th order LP coefficients a and associated gain g_{LP} were calculated, the pitch ω_0 was determined using the pitch detector of [10], and the degree of voicing v was estimated using the probability of voicing measure described in [1].

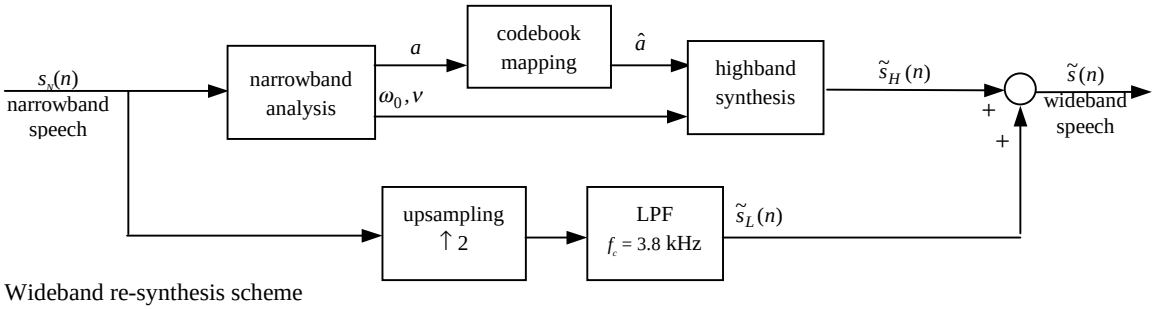

Fig. 1 Wideband re-synthesis scheme

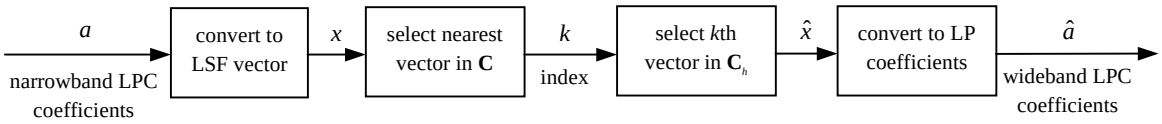

Fig. 2 Block diagram of codebook mapping scheme

3.3 Codebook training

The approach adopted here is similar to that of [3], and relies on training a pair of codebooks in parallel. Around 30 minutes of speech data produced by many different male and female speakers was obtained from the TIMIT database and used to form 87000 frames of wideband (8 kHz bandlimited) training data. This was then low pass filtered (with cutoff frequency 4 kHz) to produce 87000 frames of narrowband training data. The line spectral frequencies (LSFs) x for each frame of the narrowband training data were calculated from the 10th order LP coefficients to form a narrowband training codebook **T**. 12th order spectral linear prediction [8] was applied to the 3-8 kHz portion of the STFT of each frame of wideband speech, and the resulting coefficients were converted to LSFs $\hat{x}$, forming a highband training codebook $\mathbf{T}_h$. An LBG algorithm [7], modified for wideband extension, was then applied to optimize these codebooks (**T**, $\mathbf{T}_h$), producing a pair of codebooks (**C**, $\mathbf{C}_h$) of size 1024.

3.4 Codebook mapping

Having formed the required codebooks, the mapping procedure was straightforward, as shown in Fig. 2. The highband short-term spectral envelope was determined by finding the index k of the narrowband code vector with the closest LSF values (unweighted Euclidean distance) to those of the narrowband input speech frame, and then calculating the spectral envelope from the LSFs contained in the highband code vector $\hat{x}$ with index k.

3.5 Highband spectral modelling and gain calculation

Two important considerations in this work are accurately modelling the highband spectral envelope and matching the gain of the highband envelope to that of the narrowband envelope. [3] employs a wideband (0-8 kHz) envelope with LP order 16 to model the highband spectral envelope, and a

spectral distance measure based upon the 0-4 kHz region of narrowband and wideband envelopes. This approach causes 8 poles to be arbitrarily spread over the entire wideband, and it was found experimentally that using a highband envelope (3-8 kHz) in place of a wideband envelope improved the modelling of the high band. In our approach, a gain g_{SD} was calculated as the average spectral distance (dB) between the envelopes in the 3-3.5 kHz region. The total gain g_T was calculated as the product of g_{LP} and g_{SD}. Note that only the 3.8-8 kHz portion of the highband envelope was actually used in highband synthesis.

3.6 Highband synthesis

Synthesis of highband components was achieved using a mixed excitation model which assumes that highband speech is composed of both periodic and random components across the entire band (unlike the MBE-based model of [4]).

Periodic component

The periodic component of the excitation synthesis scheme employed in this work was based upon the harmonic sinusoidal model of [11], and models highband speech as

$$\hat{s}_h(n) = \sum_{m=M_N+1}^{M_W} A(m\omega_0)\exp\left[j\left(nm\omega_0 + m\phi_0 + \Phi_s(m\omega_0) + (1-v)\varepsilon\right)\right] \quad (1)$$

where $A(\omega)$ is the amplitude, ω_0 is the fundamental frequency, $\Phi_s(\omega)$ is the vocal tract system phase, ϕ_0 is the fundamental phase, M_N and M_W are the number of harmonics in the narrowband and wideband respectively, $v \in [0, 1]$ is the degree of voicing (0 = unvoiced, 1 = voiced), and $\varepsilon \in [-\pi,\pi]$ is a uniformly distributed random variable. The amplitudes $A(m\omega_0)$ and the vocal tract system phases $\Phi_s(m\omega_0)$ were evaluated using the 12th order LP coefficients $\hat{a}$ of the wideband spectral envelope and gain g_T, and the fundamental phase for the kth frame was calculated as

$$\phi_0^k = \phi_0^{k-1} + \frac{(\omega_0^{k-1} + \omega_0^k)T}{2} \quad (2)$$

where T is the frame length in seconds. Parameter interpolation between frames was achieved using the frequency tracking technique of [6], linear interpolation of the sinusoid amplitudes and cubic phase interpolation [11].

Random component

Unvoiced components in the highband were modelled using a combination of the randomised harmonic phases (1) and random excitation shaped by the spectral envelope (3). The extent of each was determined by v, the narrowband degree of voicing. The spectrum of the random excitation component for each frame of speech was synthesized as

$$\hat{S}_r(\omega) = (1-v)G\left|H_{LP}(e^{j\omega})\right|W(\omega) \quad (3)$$

where G is a gain factor (to calibrate the energy of $\hat{S}_r(\omega)$ relative to $S_h(\omega)$), $H_{LP}(e^{j\omega})$ is the transfer function of the LP filter with denominator coefficients $\hat{a}$, and $W(\omega)$ is the spectrum of a bandpass (3.8-8 kHz) uniformly distributed random sequence. The random excitation component $\ddot{s}_r(n)$ was calculated as the inverse DFT of $S_r(\omega)$, and the output highband speech was synthesized as the sum of the harmonic and random components:

$$\tilde{s}_H(n) = \hat{s}_h(n) + \hat{s}_r(n) \quad (4)$$

4. RESULTS

4.1 Closed testing

The above algorithm was first simulated in non-real time under closed testing conditions (codebooks derived directly from the speech segment under analysis). The resulting speech was perceptually indistinguishable from the original wideband speech. Experiments using harmonic-only and random-only excitation confirmed that the mixed excitation synthesis model described in section 3.6 provides a superior model for highband speech.

4.2 Open testing

A wideband speech segment which was not part of the speech data used to train the codebooks was bandlimited to 4 kHz and applied as input to the above algorithm. The re-synthesized wideband speech was compared with the narrowband input speech in informal listening tests, and was considered far superior in quality. Compared with the original wideband speech, the re-synthesized speech had barely noticeable perceptual differences.

The spectral distortion in the highband envelope over K frames was calculated as

$$D = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\frac{4}{\omega_s}\int_{\frac{\omega_s}{4}}^{\frac{\omega_s}{2}}\left[20\log_{10}\left(\frac{S_k(\omega)}{\tilde{S}_k(\omega)}\right)\right]^2 d\omega} \quad (5)$$

where $S_k(\omega)$ and $\tilde{S}_k(\omega)$ are magnitude spectra of the k'th (temporally aligned) frames of the original and reconstructed wideband speech respectively and ω_s is the sampling frequency. Over an arbitrarily selected 5 second speech segment, an average highband spectral distortion of 8.6 dB was measured.

Distortion in the highband envelopes under open testing conditions was predominantly due to the quantization effect inherent in codebook mapping. Larger codebooks could be employed to reduce this distortion, subject to memory implementation constraints. While the distortion obtained in this work was small, it was nevertheless perceptually significant and suggests the need for further investigation of envelope mapping methods. Assessment of this scheme using formal subjective listening tests and comparison with existing

wideband coding standards will be required to ascertain its full potential.

5. CONCLUSION

A speech enhancement scheme capable of extending the bandwidth of narrowband speech to produce high quality wideband speech has been proposed. This scheme demonstrates potential either as a post-processor to any standard narrowband coder or for incorporation into a low bit rate wideband coder. A mixed highband excitation model, realized using sinusoidal synthesis combined with spectrally-shaped random excitation, produced wideband speech which was perceptually indistinguishable from the original wideband speech. Together with an appropriate envelope mapping method such as codebook mapping, this scheme is capable of generating high quality wideband speech from narrowband speech produced by any speaker.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the assistance of Motorola Australia, and wish to express their thanks to Dr M. Thomson for his advice and encouragement.

REFERENCES

[1] Atal, B. S., and Rabiner, L. R., "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition", *IEEE Trans. Acoust., Sp., and Sig. Proc.*, vol. ASSP-24, no. 23, June 1976, pp. 201-212.

[2] Avendano, C., Hermansky, H., and Wan, E. A., "Beyond Nyquist: towards the recovery of broad-bandwidth speech from narrow-bandwidth speech", in *Proc. 4th European Conf. On Speech Commun. and Technol., EUROSPEECH, Madrid*, Sept. 1995, vol. 1, pp. 165-168.

[3] Carl, H., and Heute, U., "Bandwidth enhancement of narrowband speech signals", *SIGNAL PROCESSING VII, Theories and Applications, EUSIPCO*, 1994, vol. 2, pp. 1178-1181.

[4] Chan, C-F., and Hui, W-K., "Wideband enhancement of narrowband coded speech using MBE re-synthesis", *Proc. Int. Conf. on Signal Processing, ICSP*, 1996, vol. 1, pp. 667-670.

[5] Cheng, Y. M., O'Shaughnessy, D., and Mermelstein, P., "Statistical recovery of wideband speech from narrowband speech", *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, October 1994, pp. 544-548.

[6] Griffin, D. W., and Lim, J. S., "Multiband excitation vocoder", *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 8, August 1988, pp. 1223-1235.

[7] Linde, Y., Buzo, A., and Gray, R. M., "An algorithm for vector quantizer design", *IEEE Trans. Commun.*, vol. COM-28, no. 1, January 1980, pp. 84-95.

[8] Makhoul, J., "Spectral linear prediction: properties and applications", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-23, no. 3, June 1975, pp. 283-296.

[9] Makhoul, J., and Berouti, M., "High frequency regeneration in speech coding systems", in *Proc. Int. Conf. Acoust., Speech, Signal Processing, ICASSP*, Washington D.C., USA, 1979, pp. 428-431.

[10] McAulay, R. J., and Quatieri, T. F., "Pitch estimation and voicing detection based on a sinusoidal speech model", in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 1990, pp. 249-252.

[11] McAulay, R. J., and Quatieri, T. F., "Sinusoidal coding", in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal (Eds), Elsevier, Amsterdam,1995, Chapter 4, pp. 121-173.

[12] Nakatoh, Y., Tsushima, M., and Norimatsu, T., "Generation of broadband speech from narrowband speech using piecewise linear mapping", in *Proc. 5th European Conf. on Speech Commun. and Technol., EUROSPEECH, Rhodes*, Sept. 1997, vol.3, pp. 1643-1646.

[13] Patrick, P. J., and Xydeas, C. S., "Speech quality enhancement by high frequency band generation", *Digital processing of signals in communications:* Loughborough, 7th-10th April, 1981, pp. 365-373 (*Proc IERE*; no. 49).

[14] Paulus, J. W., and Schnitzler, J., "16 kbit/s wideband speech coding based on unequal subbands", in *Proc. Int. Conf. Acoust., Speech, Signal Processing, ICASSP*, Atlanta, Georgia, USA, 1996, pp. 157-160.

[15] Seymour, C. W., and Robinson, A. J., "A low-bit-rate speech coder using adaptive line spectral frequency prediction", in *Proc. 5th European Conf. On Speech Commun. and Technol., EUROSPEECH, Rhodes*, Sept. 1997, vol. 3, pp. 1319-1322.

[16] Yasukawa, H., "Spectrum broadening of telephone band signals using multirate processing for speech quality enhancement", *IEICE Trans. Fundamentals*, vol. E78-A, no. 8, August 1995, pp. 996-998.

[17] Yoshida, Y., and Abe, M,. "An algorithm to reconstruct wideband speech from narrowband speech based on codebook mapping", in *Proc. Int. Conf. on Spoken Language Processing, ICSLP, Yokohama*, 1994, pp. 1591-1594.