VOICE SOURCE ESTIMATION FOR ARTIFICIAL BANDWIDTH EXTENSION OF TELEPHONE SPEECH

Mark R. P. Thomas, Jon Gudnason, Patrick A. Naylor

Communications and Signal Processing Group
Imperial College London
Exhibition Road, London SW7 2AZ, UK

{mrt102|jg|p.naylor}@imperial.ac.uk

Bernd Geiser, Peter Vary

Institute of Communication Systems
and Data Processing (ind)
RWTH Aachen University, Germany

{geiser|vary}@ind.rwth-aachen.de

ABSTRACT

Artificial bandwidth extension (ABWE) of speech signals aims to estimate wideband speech (50 Hz – 7 kHz) from narrowband signals (300 Hz – 3.4 kHz). Applying the source-filter model of speech, many existing algorithms estimate vocal tract filter parameters independently of the source signal. However, many current methods for extending the narrowband voice source signal are limited to straightforward signal processing techniques which are only effective for high-band estimation. This paper presents a method for ABWE that employs novel data-driven modelling and an existing spectral mirroring technique to estimate the wideband source signal in both the high and low extension bands. A state-of-the-art Hidden Markov Model-based estimator evaluates the temporal and spectral envelopes in the missing frequency bands, with which the ABWE speech signal is synthesized. Informal listening tests comparing two existing source estimation techniques and two permutations of the proposed approach show an improvement in the perceived bandwidth of speech signals, in particular towards low frequencies. Subjective tests on the same data show a preference for the proposed techniques over the existing methods under test.

Index Terms— Speech enhancement, artificial bandwidth extension, voice source modelling

1. INTRODUCTION

The audio bandwidth of 300 Hz – 3.4 kHz which is used in today's fixed and mobile communication systems is comparable to that of early-day analogue telephony. When digital standards were first established, a common audio bandwidth facilitated interoperability between the analogue and digital domains. There has since been motivation within the telecommunications industry to introduce wideband telephony which can deliver high-quality speech with an audio bandwidth of 50 Hz – 7 kHz to end-user terminals. However, both narrowband and wideband systems are expected to co-exist for a long time, requiring measures to ensure interoperability between narrowband and wideband telephones.

This coexistence poses two main challenges: (a) efficient transcoding between narrowband and wideband signals, and (b) speech bandwidth extension to improve the quality of narrowband speech received on wideband terminals. The former can been addressed by hierarchical coding where a standard narrowband bitstream is augmented with side information to extend the audio bandwidth [1]. This approach is termed *bandwidth extension with side information*. Transcoding is then straightforward as the side information be either included or discarded as required [1]. In the latter case, the so-called extension bands (50 – 300 Hz and 3.4 – 7 kHz) are instead *estimated* from the narrowband speech only. This is referred to as *Artificial Bandwidth Extension* (ABWE).

Most ABWE methods use the source-filter model of speech production to estimate wideband spectral and temporal envelopes independently of the source signal. Appropriate techniques to blindly estimate these envelopes include codebook mapping [2], piece-wise linear mapping [3] and Bayesian methods based on Gaussian Mixture Models (GMMs) [4] or Hidden Markov Models (HMMs) [5]. Although the existing methods can already deliver improved audio bandwith compared to narrowband speech, many ABWE algorithms employ relatively crude methods to extend the source signal. For ABWE towards high frequencies (3.4 – 7 kHz) there is evidence that the quality of the enhanced speech mainly depends on a precise estimate of the spectral envelope while the source signal extension is less important [6]. However, if low audio frequencies (50 – 300 Hz) are also to be recovered from narrowband speech, existing source extension methods usually fail to produce a signal of sufficient quality, in particular for voiced speech segments. Typical artefacts include a roughness caused by low-frequency random noise that is modulated by the speech amplitude, or a buzziness caused by incorrectly shaped or incorrectly placed glottal pulses, depending upon the method employed. Such artefacts render the bandwidth-extended speech unnatural and can mask any perceived improvement in speech quality. For this reason, existing ABWE approaches often avoid lowband extension altogether.

This paper presents a novel method for the extension of narrowband source signals based on an existing spectral mirroring technique and Data-Driven Voice Source Modelling (DDVSM) [7], employing GMMs to establish an explicit mapping between narrowband source features and the wideband source signal. Using an existing ABWE framework [5] that applies HMM-based Bayesian estimation of spectral and temporal envelopes [1], missing frequency content in both high and low bands is synthesized and added to the narrowband signal to form an estimated wideband signal. Informal listening tests show that this approach achieves a particular improvement in the lowband speech signal. Subjective testing demonstrates that a noticeable improvement in the speech bandwidth is perceived at the expense of introducing some unwanted artefacts.

The remainder of this paper is organized as follows. In Section 2, existing ABWE source methods are reviewed, followed by a description of the proposed data-driven voice source technique. Section 3 introduces the estimation technique to estimate temporal and spectral envelopes. The system is evaluated in Section 4 and conclusions are drawn in Section 5.

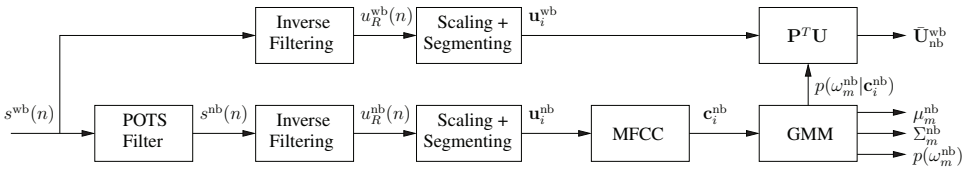

Fig. 1. System diagram for training of the proposed excitation source signal estimator.

2. VOICE SOURCE ESTIMATION

2.1. Existing Source Estimation Techniques

Several methods exist for the artificial bandwidth extension of the high band source signal. Spectral approaches involve translating, mirroring (folding), or modulating the estimated narrowband linear prediction residual, $e^{nb}(n)$ [5, 8]. Techniques that involve filtering and modulating random noise are also employed. Synthetic glottal pulses inserted in synchrony with the long-term predictor in narrowband CODECs can be used in addition to shaped noise [1].

Low band extension techniques include the generation of pitch-synchronous sinusoids [9] or impulse trains [10] and nonlinear processing of $e^{nb}(n)$ to generate low frequency harmonics with a suitable temporal envelope [11]. Such techniques are generally limited to voiced speech as unvoiced speech contains little energy below 300 Hz. Artefacts associated with lowband extension include buzzing from poorly placed or poorly shaped glottal pulses and roughness caused by incorrectly shaping additive noise.

Existing techniques make little or no use of voice source modelling. The remainder of this paper describes an entirely model-based approach for the bandwidth extension of voiced speech.

2.2. Model-Based Source Extension

2.2.1. Introduction to Data-Driven Voice Source Modelling

Data-Driven Voice Source Modelling (DDVSM) [7] is a technique for classifying voice source signals. One such implementation uses a large database of training data to estimate class distributions in the MFCC feature space, from which a set of corresponding 'prototype' time-domain waveforms are derived. An unknown voice source can then be decomposed into a weighted sum of prototypes. The technique has been modified to use a mixture of wideband and narrowband training data for the purposes of ABWE.

2.2.2. Model Training

Consider a frame of wideband speech, $s^{\mathrm{wb}}(n)$, with z-transform $S^{\mathrm{wb}}(z)$ such that

$$S^{\mathrm{wb}}(z) = U^{\mathrm{wb}}(z)V^{\mathrm{wb}}(z)R(z) = U_R^{\mathrm{wb}}(z)V(z), \qquad (1)$$

where $U^{\mathrm{wb}}(z)$ is glottal volume velocity and $R(z)$ is a model of lip radiation. The linearity of the decomposition permits the first and last terms to be encompassed into a single 'source' signal, $U_R^{\mathrm{wb}}(z)$, that excites the vocal tract filter to produce speech. A p-th order linear predictor yields an all-pole estimate of the vocal tract filter, $\hat{V}^{\mathrm{wb}}(z) \simeq V^{\mathrm{wb}}(z)$. Inverse-filtering $S^{\mathrm{wb}}(z)$ with $\hat{V}^{\mathrm{wb}}(z)$ estimates the source signal,

$$\frac{S^{\mathrm{wb}}(z)}{\hat{V}^{\mathrm{wb}}(z)} \simeq U_R^{\mathrm{wb}}(z) \leftrightharpoons u_R^{\mathrm{wb}}(n). \qquad (2)$$

Let $s^{\mathrm{nb}}(n)$ be the corresponding narrowband speech signal, with voice source $u_R^{\mathrm{nb}}(n)$, obtained with a plain-old telephone system (POTS) filter whose passband lies in 300 Hz – 3.4 kHz.

The APLAWD database [12], which contains wideband speech signals and contemporaneous Electroglottogram (EGG) recordings, forms the training corpus for the voice source model training. The SIGMA algorithm [13] detects glottal closure instants (GCIs) from the EGG signal, which are then refined by finding the maximum gradient in $u_R^{\mathrm{wb}}(n)$ that lies ± 0.5 ms of each SIGMA-derived GCI. This corrects for small deviations in the EGG-to-speech time alignment. The estimated source signals are divided into scale- and amplitude-normalized overlapping two-cycle glottal-synchronous frames so that classification is based only on waveform shape,

$$\begin{aligned} \mathbf{u}_i^{\mathrm{wb}} &= \updownarrow_L^{L'} \kappa u_R^{\mathrm{wb}}(n), \\ \mathbf{u}_i^{\mathrm{nb}} &= \updownarrow_L^{L'} \kappa u_R^{\mathrm{nb}}(n), n \in \{n_i^c, \ldots, n_{i+2}^c - 1\}, \end{aligned} \qquad (3)$$

where $\updownarrow_L^{L'}$ denotes a resampling of factor $\frac{L'}{L}$, $L = n_{i+2}^c - n_i^c + 1$, $L' = 2t_{\max}f_s$, n_i^c is the GCI at cycle i, t_{max} is a maximum glottal period of 0.02 ms, f_s is sampling frequency (Hz) and κ is a gain factor to normalize RMS energy. Cycle pairs form the rows of $(N \times L')$ data matrices where N is the total number of cycle pairs,

$$\begin{aligned} \mathbf{U}^{\mathrm{wb}} &= [\mathbf{u}_1^{\mathrm{wb}}, \mathbf{u}_2^{\mathrm{wb}}, \ldots, \mathbf{u}_N^{\mathrm{wb}}]^{\mathrm{T}}, \\ \mathbf{U}^{\mathrm{nb}} &= [\mathbf{u}_1^{\mathrm{nb}}, \mathbf{u}_2^{\mathrm{nb}}, \ldots, \mathbf{u}_N^{\mathrm{nb}}]^{\mathrm{T}}. \end{aligned} \qquad (4)$$

A $(N \times C)$ feature matrix of $C = 12$ MFCCs is derived for each narrowband frame,

$$\mathbf{C}^{\mathrm{nb}} = [\mathbf{c}_1^{\mathrm{nb}}, \mathbf{c}_2^{\mathrm{nb}}, \ldots, \mathbf{c}_N^{\mathrm{nb}}]^{\mathrm{T}}, \qquad (5)$$

from which the EM algorithm [14] derives $M = 16$ diagonal covariance Gaussian mixtures. The probability that feature $\mathbf{c}_i^{\mathrm{nb}}$ is a member of mixture component ω_m is stored as an $(N \times M)$ probability matrix with elements $p(\omega_m^{\mathrm{nb}}|\mathbf{c}_i^{\mathrm{nb}})$. For each mixture, the corresponding class centroids, μ_m^{nb}, diagonal covariance matrices, Σ_m^{nb} and mixture weights, $p(\omega_m^{\mathrm{nb}})$, are calculated. The prototype signals are derived as a weighted average of wideband time-domain waveforms, $\mathbf{u}_i^{\mathrm{wb}}$, stored in a $(M \times L')$ matrix,

$$\bar{\mathbf{U}}_{\mathrm{nb}}^{\mathrm{wb}} = [\bar{\mathbf{u}}_{\mathrm{nb},1}^{\mathrm{wb}}, \bar{\mathbf{u}}_{\mathrm{nb},2}^{\mathrm{wb}}, \ldots, \bar{\mathbf{u}}_{\mathrm{nb},M}^{\mathrm{wb}}]^{\mathrm{T}} = \mathbf{P}_{\mathrm{nb}}^{\mathrm{T}}\mathbf{U}^{\mathrm{wb}}. \qquad (6)$$

We employ a convention for $\mathbf{U}$ and $\mathbf{P}$ whereby the superscript refers to the bandwidth of the time-domain waveforms and the subscript to that of the feature set. The system is depicted in Fig. 1.

2.2.3. Wideband Voice Source Estimation

Wideband voice source estimation is similar to model training. A narrowband test utterance is inverse-filtered and segmented into amplitude and scale-normalized 2-cycle frames, $\mathbf{u}_i^{\mathrm{nb}}$, with corresponding MFCCs, $\mathbf{c}_i^{\mathrm{nb}}$. The DYPSA algorithm [15] provides estimation of GCIs for segmentation. The decomposition for frame i is

$$\boldsymbol{\gamma}_i = [\gamma_{1,i}, \gamma_{2,i}, \ldots, \gamma_{M,i}] = [p(\omega_1^{\mathrm{nb}}|\mathbf{c}_i^{\mathrm{nb}}), \ldots, p(\omega_M^{\mathrm{nb}}|\mathbf{u}_i^{\mathrm{nb}})]. \qquad (7)$$

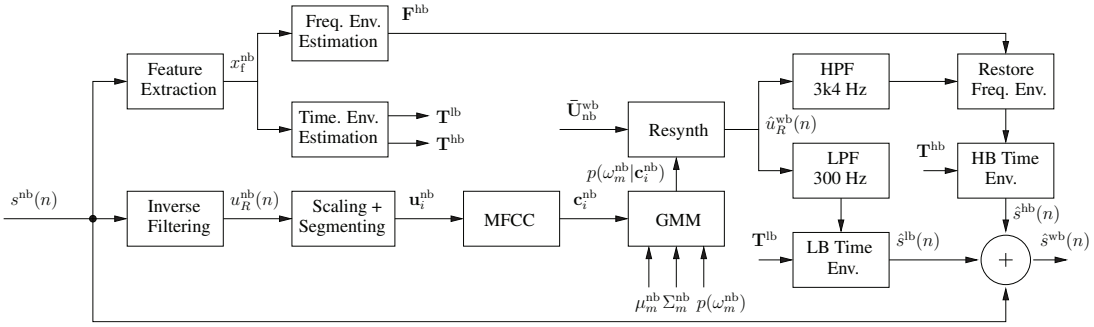

Fig. 2. BWE synthesis.

We define a set $\mathcal{M}_i \subseteq \mathcal{M}_{\mathrm{all}}$, $\mathcal{M}_{\mathrm{all}} = \{1, \ldots, M\}$. $\mathcal{M}_i$ contains the class indices that produce the highest likelihood. A wideband cycle of the voice source signal can then be resynthesized from the prototypes, $\bar{\mathbf{U}}_{\mathrm{nb}}^{\mathrm{wb}}$, with the decomposition terms,

$$\hat{\mathbf{u}}_i^{\mathrm{wb}} = \sum_{m \in \mathcal{M}_i} \gamma_{m,i} \left(\updownarrow_{\alpha}^{\beta} \kappa \bar{\mathbf{u}}_{\mathrm{nb},m}^{\mathrm{wb}} \right), \tag{8}$$

where $\updownarrow_{\alpha}^{\beta}$ resamples $\bar{\mathbf{u}}_m$ to length $(n_{i+2}^c - n_i^c)$ for cycle i and κ is a gain factor to reproduce the same energy as the source cycle. An approximation to the full $u_{\mathrm{R}}(n)$ is synthesized by windowing $\hat{\mathbf{u}}_i$ with a Hamming window, $\mathbf{w}_i$, shifting to centre on n_i^c and summing.

The proposed approach is suitable for voiced speech only; unvoiced excitation can be produced with a spectral technique such as mirroring. A voiced/unvoiced/silence detector [16] is employed.

3. TEMPORAL & SPECTRAL ENVELOPE ESTIMATION

To complete the ABWE scheme, the wideband signal envelope has to be estimated and restored. Several envelope parameterizations have been proposed for ABWE. Mostly, an autoregressive model is assumed and the envelope is restored using an LPC synthesis filter with estimated coefficients. However, LPC synthesis of artificial source signals does not necessarily regenerate the correct temporal characteristics. Therefore, in this work, a signal parameterization is employed in terms of spectral *and* temporal energy envelopes [1], whereby low and high extension bands are treated separately. For the high extension band, the spectral envelope is parameterized in terms of 10 logarithmic subband energies, $\mathbf{F}_{\mathrm{hb}}$, (375 Hz subbands) for each 10 ms frame. The temporal envelope, $\mathbf{T}_{\mathrm{hb}}$, provides 5 logarithmic subframe energies of the extension band signal for each 10 ms frame (2 ms subframes). For the low extension band, only the temporal envelope, $\mathbf{T}_{\mathrm{lb}}$, of the low-pass signal is used.

The parameter vectors $\mathbf{F}_{\mathrm{hb}}$, $\mathbf{T}_{\mathrm{hb}}$ and $\mathbf{T}_{\mathrm{lb}}$ are estimated with separate HMM-based MMSE estimators [5] The estimators require a narrowband feature vector, $x_{\mathrm{f}}^{\mathrm{nb}}$, for each frame. Here, $x_{\mathrm{f}}^{\mathrm{nb}}$ is composed of the narrowband MFCCs, of the zero crossing rate and of the narrowband temporal envelope $\mathbf{T}_{\mathrm{nb}}$. The actual estimator configurations are listed in Table. 1. Based on the estimated parameter set, the extension band signals $\hat{s}^{\mathrm{hb}}(n)$ and $\hat{s}^{\mathrm{lb}}(n)$ can be synthesised by shaping the envelopes of the source signals, $\hat{u}^{\mathrm{hb}}(n)$ and $\hat{u}^{\mathrm{lb}}(n)$, respectively. This signal shaping is performed in a two-step approach: a filterbank equalizer restores the spectral envelope (high band only) and the temporal envelope is corrected via gain manipulation, cf. [1]. Finally, the signals $\hat{s}^{\mathrm{hb}}(n)$ and $\hat{s}^{\mathrm{lb}}(n)$ are combined with $s^{\mathrm{nb}}(n)$ to give the bandwidth extended output $\hat{s}^{\mathrm{wb}}(n)$. The estimation / resynthesis procedure is shown in Fig. 2.

4. EVALUATION

Four voice source estimation techniques were considered for subjective testing: i) spectral mirroring, ii) synthetic glottal pulse located at the GCIs during voiced + spectral mirroring during unvoiced, iii) DDVSM during voiced + spectral mirroring during unvoiced and iv) DDVSM for LB + spectral mirroring for HB. The ABWE techniques were applied to narrowband speech, quantized with an ITU-T G.711 μ-law audio CODEC [17].

An ITU-T P.800 [18] subjective test was devised, including two additional hidden references in the form of wideband and quantized narrowband speech. The sample set consisted of 3 female and 3 male talkers, each speaking 5 pairs of phonetically-balanced sentences. The 20 subjects each listened to the 30 samples in random order, with one of the 6 methods randomly applied to each sentence. Processed sentences were normalized to a level of -30 dB with respect to the overload point defined in ITU-T P.56 [19], then presented with Sennheiser HD650 headphones in a listening room environment. Subjects were asked to rate i) 'Foreground', describing speech quality only, ii) 'Background', describing artefact tolerance, and iii) 'Overall' impression. An Absolute Category Rating (ACR) scale was used for i) and iii) and a Degradation Category Scale (DCR) for ii), rated 1 – 5 in 0.5 increments. Five examples were given with approximate ratings prior to taking the test. A set of 'control' samples, rated by a team of expert listeners, were used to derive a quadratic calibration curve for each subject to standardize their responses.

The results show that all ABWE techniques improve the perceived foreground score at the expense of reducing the background score. Of the techniques under test, a clear preference was shown for the combined DDVSM LB + spectral mirroring HB, confirming the assertion that DDVSM is particularly effective for lowband extension and that lowband ABWE is especially sensitive to the source signal employed. The preference of the best ABWE technique compared with narrowband is still relatively small. These results contrast with previous findings where highband-only ABWE is preferred to narrowband, suggesting that lowband artefacts are particularly detrimental to perceived quality.

Table 1. Envelope Estimator Configurations

Param. Vect.	Param. Dim.	# of Features	Codebook Size	# of Gaussians.
$\mathbf{F}_{\mathrm{hb}}$	10	19	128	8
$\mathbf{T}_{\mathrm{hb}}$	5	19	128	8
$\mathbf{T}_{\mathrm{lb}}$	5	19	64	8

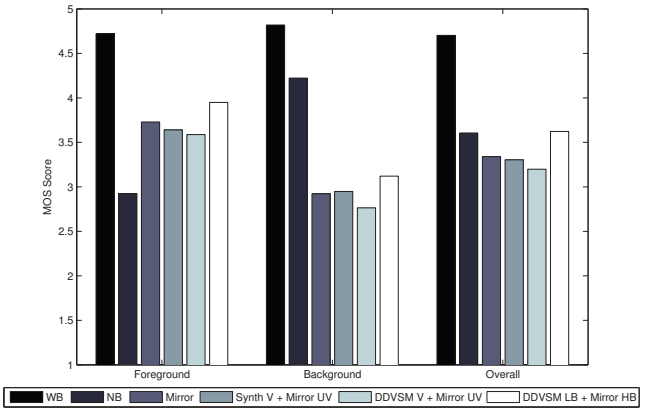

Fig. 3. Mean Opinion Scores for ABWE algorithms.

Two artefacts regularly occurred in the test set. The first was a 'beating' in the low extension band, caused by erroneous GCI detections, resulting in a low frequency excitation signal that was not pitch-synchronous with the narrowband signal. The second was a 'hissing' in the high extension band, caused by incorrect estimation of the upper spectral envelope. Improved GCI detection, coupled with fine-tuning of temporal/spectral envelope training and estimation, have resulted in significantly reduced artefacts since the initial submission of this paper.

5. CONCLUSIONS

An artificial bandwidth extension (ABWE) technique has been proposed that employs spectral mirroring and Data-Driven Voice Source Modelling (DDVSM) to estimate a wideband source signal from narrowband speech. Used in conjunction with a state-of-the-art framework that estimates the temporal and spectral envelopes of the extension bands, an ABWE system has been proposed that is novel in its explicit use of voice source modelling and the estimation both the low (50 – 300 Hz) the high (3.4 – 7 kHz) extension bands.

Informal listening tests reveal that the proposed technique is particularly effective in the lowband. Formal subjective tests demonstrate that an improvement in the perceived bandwidth of speech can be achieved at the expense of increasing background artefacts. It further reveals that, compared with the other methods under test, the use of DDVSM in the lowband with spectral mirroring in the highband is preferred over narrowband speech, as it provides the greatest perceived ABWE with the least number of unwanted artefacts.

6. REFERENCES

[1] B. Geiser et al., "Bandwidth extension for hierarchical speech and audio coding in ITU-T Rec. G.729.1," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2496–2509, Nov. 2007.

[2] H. Carl and U. Heute, "Bandwidth enhancement of narrowband speech signals," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Edinburgh, Scotland, Sept. 1994, pp. 1178–1181.

[3] Y. Nakatoh, M. Tushima, and T. Norimatsu, "Generation of broadband speech from narrowband speech using piecewise linear mapping," in *Proc. European Conf. on Speech Communication and Technology*, Rhodos, Greece, Sept. 1997, vol. 3, pp. 1643–1646.

[4] Kun-Youl Park and Hyung Soon Kim, "Narrowband to wideband conversion of speech using GMM based transformations," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, June 2000, vol. 3, pp. 1843–1846.

[5] Peter Jax and Peter Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, Aug. 2003.

[6] Hannu Pulakka, Paavo Alku, Laura Laaksonen, and Päivi Valve, "The effect of highband harmonic structure in the artificial bandwidth expansion of telephone speech," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 2497–2500.

[7] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Data-driven voice source waveform modelling," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009.

[8] J. A. Fuemmeler, R. C. Hardie, and W. R. Gardner, "Techniques for the regeneration of wideband speech from narrowband speech," *EURASIP Journal on Applied Signal Processing*, , no. 4, pp. 266–274, Dec. 2001.

[9] Jean-Marc Valin and Roch Lefebvre, "Bandwidth extension of narrowband speech for low bit-rate wideband coding," in *Proc. IEEE Speech Coding Workshop*, Delavan, WI, USA, Sept. 2000, pp. 130–132.

[10] Ismail Uysal, Harsha Sathyendra, and John G. Harris, "Bandwidth extension of telephone speech using frame-based excitation and robust features," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Antalya, Turkey, Sept. 2005.

[11] U. Kornagel, "Spectral widening of the excitation signal for telephone-band speech enhancement," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Darmstadt, Germany, Sept. 2001, pp. 215–218.

[12] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," Technical report, University College London, June 1987.

[13] M. R. P. Thomas and P. A. Naylor, "The SIGMA algorithm: A glottal activity detector for electroglottographic signals," *IEEE Trans. Audio, Speech, Lang. Process.*, 2009, to appear.

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

[15] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Speech Audio Process.*, vol. 15, no. 1, pp. 34–43, Jan. 2007.

[16] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Application of the DYPSA algorithm to segmented time-scale modification of speech," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008.

[17] ITU-T, "Pulse code modulation (PCM) of voice frequencies," Nov. 1998.

[18] "Methods for subjective determination of transmission quality," Aug. 1996.

[19] ITU-T, "Objective measurement of active speech level," Mar. 1993.