Towards a New Image-Based Spectrogram Segmentation Speech Coder Optimised for Intelligibility

K.A. Jellyman[1], N.W.D. Evans[1,2], W.M. Liu[1], and J.S.D. Mason[1]

[1] School of Engineering, Swansea University, UK
174869@swan.ac.uk, 199997@swan.ac.uk, j.s.d.mason@swan.ac.uk
[2] EURECOM, Sophia Antipolis, France
nicholas.evans@eurecom.fr

Abstract. Speech intelligibility is the very essence of communications. When high noise can degrade a speech signal to the threshold of intelligibility, for example in mobile and military applications, introducing further degradation by a speech coder could prove critical. This paper investigates concepts towards a new speech coder that draws upon the field of image processing in a new multimedia approach. The coder is based on a spectrogram segmentation image processing procedure. The design criterion is for minimal intelligibility loss in high noise, as opposed to the conventional quality criterion, and the bit rate must be reasonable. First phase intelligibility listening test results assessing its potential alongside six standard coders are reported. Experimental results show the robustness of the LD-CELP coder, and the potential of the new coder with particularly good results in car noise conditions below -4.0dB.

1 Introduction

Speech communications has been revolutionised by mobile communications allowing phone calls to be made almost "anywhere and at anytime" [1]. One consequence of this expectation is the increased potential for background noise that can be sufficiently strong so that it threatens intelligibility.

Intelligibility is the very essence without which communication does not exist. Originally high levels of noise would have been more common place with military and security applications. Now though, phone usage is no longer restricted to the typical relatively quiet home and office environments. It is therefore perhaps surprising that despite the potential for high levels of background noise relatively little attention has been given to the topic of intelligibility assessment, certainly when compared with the more general overarching speech quality assessment [2]. It is perhaps even more surprising that this is the case even in military and security applications [3, 4, 5, 6].

Quality is all encompassing and includes intelligibility along with many other attributes including naturalness, ease of listening, and loudness. Unfortunately predicting intelligibility from overall quality tends not to be straightforward. A number of authors have observed this including [3, 2, 7, 8].

B. Huet et al. (Eds.): MMM 2009, LNCS 5371, pp. 63–73, 2009.

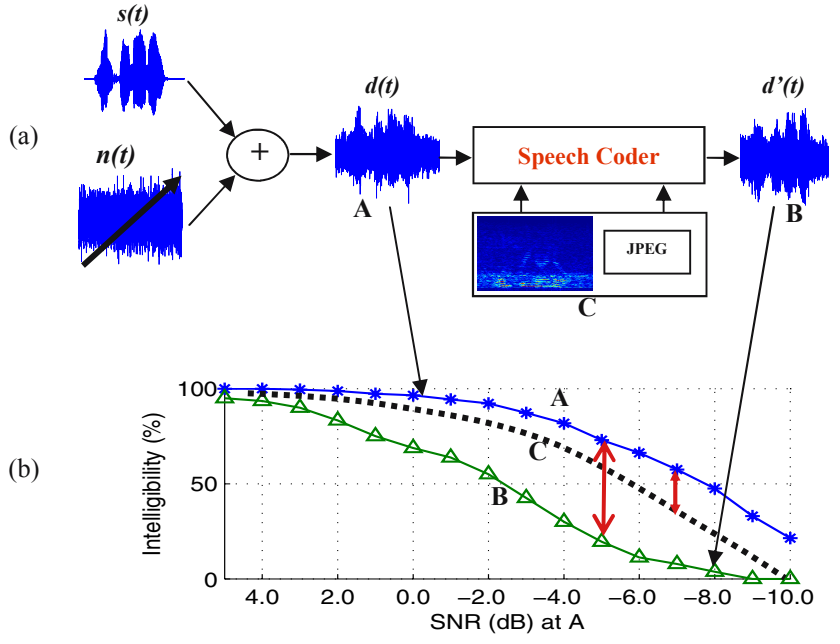

Fig. 1. Figure (a) shows clean speech degraded by an increasing level of additive noise, (A), followed by encoding and de-coding, to give speech further degraded by the coding (B) with the corresponding intelligibility profile (B) in Figure (b); the coder profile (B) is a standard MELP coder. The goal is to reduce this additional degradation by using an alternative coder, to give enhanced (here hypothetical) profile, example (C). A new image-based spectrogram segmentation coder that uses JPEG compression is investigated for this role.

The situation considered is illustrated in Figure 1(a). A clean speech signal $s(t)$ is combined with high levels of additive noise $n(t)$. As the SNR decreases the intelligibility of the signal combination $d(t)$ falls, as shown in Figure 1(b), profile A. Profiles A and B in Figure 1(b) come from intelligibility tests performed by a small group of listeners. Following the coding and transmission operations it is likely that the resultant signal $d'(t)$ suffers further degradation and consequently now exhibits lower intelligibility; this is shown in the lowest profile in Figure 1(b), profile B. This profile comes from a second set of tests performed by the same small group of listeners. The contribution to the additional fall in intelligibility (profile A to profile B) is due to the encoder and decoder operations, in this case a standard low bit rate MELP coder [9]. It can be seen that the level of intelligibility loss due to the coder increases and is particularly severe when the SNR is in the region of -2 to -8dB. The goal of the work presented here is to design a speech coder which minimises this additional coder degradation, while maintaining a reasonable bit rate. For illustrative purposes this is indicated by the hypothetical coder profile, labelled C in Figure 1(a).

The major contribution of this paper is in the investigation of concepts towards a new image-based spectrogram segmentation speech coder designed for intelligibility preservation in high noise conditions. The segmentation procedure, originally proposed by Hory and Martin [10], identifies potentially useful speech dominant information in time and frequency. The coder fuses both speech and image processing techniques in a new multimedia approach applied to the well researched problem of speech coding.

In the reported experiments we consider utterances comprising of connected, four-digit strings. The utterances span typically 1.5 to 2.0s. The coder is therefore not suitable for normal telephony usage because of the inherent delay from using spectrograms. For conversational communications the delay must typically not exceed 0.3s [11]. Thus this coder is targeted towards one way communication applications, such as military and security recording systems, where delay can be readily tolerated.

The structure of this paper is as follows: Section 2 presents an assessment of 6 standard coders, with bit rates ranging from 2.4kb/s up to 32kb/s, for their contributions to intelligibility in high noise. The assessment is presented in a manner similar to that used to derive profile B in Figure 1(b). The results from the standard coders form benchmarks for the new coder. This assessment is believed to be a first comparing a range of coders under otherwise identical noise conditions using intelligibility as the cost function; Section 3 describes the experimental image-based spectrogram segmentation coder; and Section 4 presents an intelligibility assessment of the spectrogram segmentation coder in comparison to the results from Section 2.

2 Assessment of Standard Coders

Six coders are assessed here. They are: (i) the G.721 adaptive differential pulse code modulation coder [12]; (ii) the adaptive multi-rate (AMR) coder [13]; (iii) the low delay-code excited linear predictive coder [14] (iv) the Groupe Special Mobile-full rate (GSM-FR) coder [15]; (v) the mixed excitation linear predictive (MELP) coder [9]; and (vi) the linear predictive coder (LPC-10) [16]. All six coders are used widely, are reported to have reasonable speech quality performances and have bandwidths of 32.0kb/s, 12.2kb/s, 16.0kb/s, 13kb/s, 2.4kb/s and 2.4kb/s respectively. Software implementations of all six coders are freely available at [17, 18, 19, 20, 21, 22].

2.1 Intelligibility Assessment

Reliable intelligibility assessment is an extremely difficult task. Human opinion is costly, and scores can vary across vocabulary, language, context, listeners and many such practical factors. To help circumvent some of the difficulties we restrict our assessment to digit strings, following a procedure which we proposed in [23]. Digits provide for a straightforward scoring process with minimal dependence on listeners' language abilities. Whilst it is acknowledged that digits have a limited phonetic range the use of wider vocabularies would possibly lead

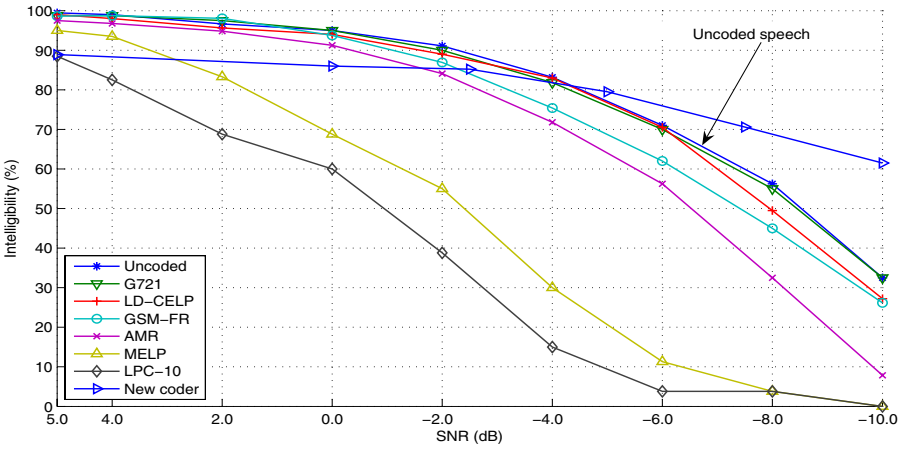

Fig. 2. Subjective intelligibility scores in car noise conditions with uncoded, with the 6 standard coders, and with the new coder. The profile for the new coder begins at 89% and crosses the uncoded profile -4.0dB. This implies that below this SNR the coder actually enhances intelligibility; however these results pertain to a fixed segmentation mask derived at 5dB SNR. Work with masks derived from the lower SNRs continues. The SNRs considered for the new coder are at 5, 0, -2.5, -5.0, -7.5 and -10.0dB.

to decreased scores across the board, with the ranking remaining largely unchanged. Thus the use of digits is seen as an acceptable compromise especially where system ranking is required rather than absolute scores.

Here we assess the intelligibility in high levels of car noise, a challenging application environment. Using standard noise addition software, from ITU-T Rec. P.56 [24], car noise was added to clean speech from 5.0dB down to -10.0dB. The speech utterances were 556 four-digit utterances sampled at 8kHz selected from the ETSI-AURORA2 digit string corpus [25].

Each SNR dataset was then processed with each of the six coders giving a total of seven conditions: 1 uncoded speech + 6 coded speech sets. Listener responses were obtained by combining previous collected responses in [26] with some newly collected responses. To maximise the potential for recruitment listeners performed tests using an on-line graphical user interface which may be viewed at *http://eeceltic.swan.ac.uk/subj*. During the tests listeners keyed in the digits they heard with intelligibility indicated by the total number of digits correctly identified.

2.2 Results

Averaged intelligibility scores for the six standard coder conditions and uncoded condition are presented in Figure 2. Included in Figure 2 is the profile for the new coder described in Section 3. The graph shows decreasing intelligibility as the SNR falls from 5.0dB to -10.0dB. For uncoded speech an intelligibility score of 100% at 5dB falls to a little over 30% at -10dB. For any given SNR four out

of the six standard coder profiles, namely excluding G721 and LD-CELP, show the additional intelligibility degradation over that from the uncoded noise alone condition. The differences between the uncoded and lower four standard coder profiles is most prominent at lower SNRs. At -4.0dB, for example, MELP and LPC-10 have intelligibility scores of approximately 30% and 15% respectively. Compared with the uncoded speech at 83%, the coders introduce additional losses of 53% and 68% respectively. However the additional loss introduced by G721 and LD-CELP is negligable compared with the uncoded speech. Other than in SNR levels lower than -6.0dB, where for example LD-CELP introduces a loss of approximately 6% at -8.0dB compared with the uncoded speech, there is no meaningful difference between the two speech coders. The performance of the G721 coder is perhaps not un-expected given the high bit rate of 32kb/s. The robustness of the LD-CELP is though somewhat surprising considering that is half the bit rate of the G721 coder at 16kb/s. The results shown in Figure 2 provide a benchmark against which the performance of the experimental coder, described in the following section, may be compared.

3 Spectrogram Segmentation Coder

In this section we present an experimental image-based spectrogram segmentation speech coder designed to preserve intelligibility in high noise conditions. The overall speech quality is not of concern; however, a reasonable bit rate.

The inspiration behind the coder comes from our previous work which applied the image-based spectrogram segmentation procedure, proposed by Hory and Martin [10], to noise robust speech recognition [27]. The coder thus combines image processing techniques with speech in a multimedia-type scenario. The coding process can be considered in 3 stages, each of which are illustrated in Figure 3. They are spectrogram segmentation, phase coding and image compression. Each of the 3 coding stages and the re-synthesis stage are now presented in turn.

3.1 Spectrogram Segmentation

The result of the procedure with one spectrogram are illustrated in Figures 3(b) and 3(c). The example speech recording corresponds to a male person speaking the digit string '139O'. Figure 3(a) illustrates the time waveform of the speech signal with added car noise at a SNR of 0dB. Immediately below, in 3(b), is the corresponding spectrogram. Regions of the spectrogram that are dominated by speech are characterised by high energy pitch harmonic lines. The spectrogram segmentation procedure can be used to extract these regions to produce a segmented magnitude spectrogram as illustrated in Figure 3(c). Thus noise dominant regions can be suppressed and removed from the encoding process which now functions only on speech dominated regions, hence the potential for preserving intelligibility. Of secondary benefit is the potential for bit rate reduction achieved through noise suppression; that should incur minimal encoding costs for noise dominated regions.

The spectrogram segmentation procedure is effectively an image processing technique. Conceptually the magnitude spectrogram is considered as a whole

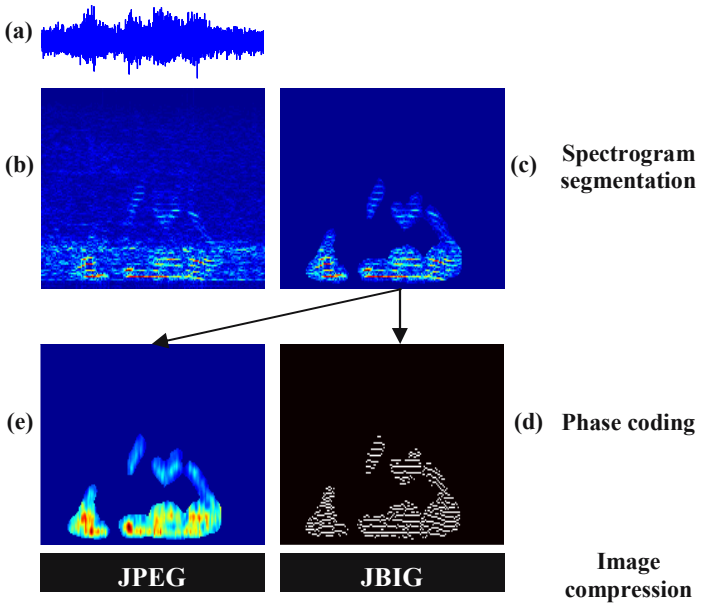

Fig. 3. Figure illustrating a 4 digit speech signal degraded by car noise to 0dB undergoing spectrogram segmentation coding

image. Features are then derived from sub-images across the whole spectrogram image. Here a sub-image size of 3 by 7 spectrograms coefficients is used that was empirically determined in [28]. The frame size is 32ms and the overlap is 8ms. The underlying principle assumed by Hory and Martin [10] is that speech and noise dominated regions are statistically different and thus identifiable. Mean and standard deviation scores for the sub-images are used as features from which a two dimensional feature space is formed. Regions dominated by either speech or noise cluster within the feature space enabling segmentation. Inherently, regions will exist in the spectrogram that essentially fall between the speech and noise dominant classes. These regions correspond to the boundaries of the dominant speech regions and thus represent regions of uncertainty.

The spectrogram segmentation procedure identifies speech dominated regions in a morphological growth process. Growth seed points in the magnitude spectrogram are selected using the feature space. The selected speech regions are then iteratively grown from the seed points. This morphological growth process continues until only noise dominant regions are deemed to remain. Hory and Martin define the end point according to the convergence of a normalised maximum likelihood [10]. A full description of the segmentation procedure is presented by Hory and Martin in [10] and also by Rodriguez *et al* in [27].

3.2 Phase Coding

Phase information as well as magnitude information is usually needed to reconstruct the time domain signal. Exploiting redundancies within the phase

spectrogram image for efficient transmission proves to be extremely difficult. The phase spectrogram appears almost entirely random with no obvious pattern. Thus in an alternate strategy we estimate the phase spectrum using the magnitude spectrogram by generating a binary peak map image. An example is shown in Figure 3(d). Peaks in the binary peak map can be seen to correspond to pitch harmonics in the segmented magnitude spectrogram in Figure 3(c). The binary peak map is generated using the principles of sinusoidal transform coding (STC) proposed by McAulay and Quatieri [29].

STC exploits the quasi-periodic nature of speech by representing speech signals as a sum of sinusoids. Each sinusoid contains three parameters that are necessary for re-synthesis, namely amplitude, frequency and phase.

$$\tilde{s}(n) = \sum_{l=1}^{L(k)} \hat{A}_l^k cos[n\hat{\omega}_l^k + \hat{\theta}_l^k] \tag{1}$$

Equation 1, taken from [29], illustrates this concept where a discretely sampled speech signal $s(n)$ estimated over a short frame k is represented by $\hat{A}_l^k$, $\hat{\omega}_l^k$ and $\hat{\theta}_l^k$ which each represent the estimated amplitude, frequency and phase for the lth sinusoid. Compression in STC is obtained by reducing the number of sinusoids needed for re-synthesis and exploiting redundancy in the 3 parameters. The relationship between magnitude and phase is highly complex. The sinusoid summation model proposed in Equation 1 effectively simplifies this problem approximating the relationship to a linear system.

The sinusoidal selection process proposed by McAulay and Quatieri [29] is a frame based peak selection process of the magnitude spectrum. In the approach adopted all the peaks over the entire frequency bandwidth in the magnitude spectrum of a frame are first selected. The peaks are then kept dependent on whether they exist in the next frame in a nearest neighbour procedure. This peak selection process leads to the "birth" and "death" concept where sinusoids start in 1 frame and end in a later frame when no continuing peaks exist in subsequent frames. An example binary peak map is shown in Figure 3(d). During voiced speech the selected sinusoids can be seen to correlate with the pitch harmonics in the original magnitude spectrogram, shown in Figure 3(b). During unvoiced speech periods McAulay and Quatieri [29] state that provided the frame increment is not more than 20ms, the sinusoidal representation is successful. Here a frame duration of 32ms and increment of 8ms is used i.e., well within the proposed limit.

The binary peak map image effectively acts as a substitute image for the phase spectrogram. Its characteristics make it far more efficient for image compression. During the re-synthesis process a random set of phase values is assigned for the first frame of a given spectrogram. The phase values are then incremented for subsequent time frames, producing the phase approximation.

To avoid redundancy and increase efficiency the pitch lines in the segmented magnitude spectrogram are removed using low pass cepstral domain filtering. The resultant smoothed spectrogram is shown in Figure 3(e).

3.3 Image Compression

The speech signal is considered as two images: a smoothed magnitude spectrogram and a binary peak map. Both must be encoded for transmission. Given that both are images we investigate the use of standard image compression techniques. Here, JPEG [30] is used to encode the smoothed magnitude and JBIG [31], a binary image encoder, for the binary peak map.

The level of image compression is variable and will influence the trade off between intelligibility and bit rate. Here bit rate is of secondary importance; the primary cost function is the maintaining of intelligibility. We have therefore chosen to set the level of JPEG compression to 20%, following initial informal intelligibility experiments to identify a knee point in the profile of intelligibility against bit rate. For the binary peak map JBIG is used in lossless mode. The corresponding combined bit rate is in the region of 17kb/s and was calculated by dividing the image file sizes for each of the 566 four digit utterances under test by their corresponding time periods. The maximum bit rate is 26kb/s and the lowest is 10kb/s.

3.4 Time Domain Re-synthesis

The re-synthesis process is in essence simply an inverse process of the image compression and spectrogram generation. The smoothed magnitude spectrogram and binary peak map are first de-compressed by reverse JPEG and JBIG coding respectively. Upon decompression the 2 images are then combined by multiplication to form 1 magnitude spectrogram image.

To reconstruct the time domain signal phase information is also needed. Following the procedure described in Section 3.2, a random set of phase values is generated for the first time frame and then advanced incrementally for subsequent frames. The inverse-discrete Fourier transform (I-DFT) is then computed for each frame to revert back to the time domain. To complete the time domain re-synthesis the framing process is then reversed by retaining only the initial frame increment period for each frame.

The time domain re-synthesis process used here represents an initial strategy which likely can be further optimised. For example, attempting to ensure smoothness at frame transition boundaries using overlap and add or interpolation procedures [29] may help to maintain intelligibility. These ideas warrant future investigation.

4 Experiments

The objective of the experiments reported here is to assess the potential of the spectrogram segmentation coder and, specifically, how well intelligibility is preserved in high noise conditions. We replace the conventional, standard speech coder, illustrated in Figure 1, with the spectrogram segmentation coder and repeat similar experiments to those described in Section 2. For the coder to

be of benefit in the current context it should operate in high noise conditions, minimising any further intelligibility loss whilst delivering a reasonable bit rate.

To assess the potential of the coder a fixed segmentation mask was used for each SNR. In each case the mask was that obtained from the same speech signal degraded at 5dB with additive white Gaussian noise (AWGN). The coder is assessed under essentially the same experimental conditions as those used for the standard coder assessment as reported in Section 2. Here, a total of 17 different listeners were used. Each SNR condition was assessed by between 7 and 10 listeners. Each listener scored a minimum of four utterances per SNR condition. An average of 43 responses were collected for each SNR condition and the listening tests were performed over a period of approximately 2 weeks.

The subjective intelligibility scores for the spectrogram segmentation coder are shown in Figure 2, combined with the earlier results from Section 2. These results show an intelligibility score of 89% at an SNR of 5.0dB, similar to the worst of the standard coders, LPC-10. However, the profile remains relatively flat down to -4.0dB at which point the coder profile coincides approximately with the no-coder condition with an intelligibility level of 82%. Below -4.0dB the coder outperforms all of the standard coders. Furthermore intelligibility scores obtained with the experimental coder exceed performance without coding. This suggests that the coder can potentially enhanced speech intelligibility provided a good segmentation mask is used.

5 Conclusion

Two contributions are made in this paper. The first is the somewhat surprising robustness found for the LD-CELP [14] coder with preserving intelligibility in high noise. The second is the investigation of concepts towards a new image based speech coder. This coder is optimised against an intelligibility criterion rather than the more common and embracing criterion of overall quality.

The coder is motivated by an image processing spectrogram segmentation procedure proposed by Hory and Martin [10]. Image processing techniques are thus fused with speech processing techniques in a new multimedia approach to speech coding. Experimental subjective intelligibility listening tests show that the coder is potentially able to enhance intelligibility in car noise levels below -4.0dB, albeit with a spectrogram segmentation mask obtained from corresponding 5dB SNR conditions. The bit rate for this coder is in the region of 17kb/s.

Work is currently on-going into developing reliable segmentation masks that are successful at lower SNRs. The segmentation procedure was developed for chirp signals degraded by AWGN [10]. A dominant characteristic of speech that is not taken advantage of in the original procedure is pitch. Two common characteristics of pitch harmonics in the magnitude spectrogram image are long lines and wide spacing between harmonics. Thus the idea being investigated is to restrict the image based segmentation procedure to narrow frequency bands. This work is on going with some early promising results.

Acknowledgements

The authors wish to thank Her Majesty's Government Communications Centre (HMGCC) for sponsoring this work.

References

1. Martin, R.: Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors. In: Proc. IEEE ICASSP, vol. 1, pp. 253–256 (2002)
2. Beerends, J.G.: Extending p.862 PESQ for assessing speech intelligibility. White contribution COM 12-C2 to ITU-T Study, Group 12 (October 2004)
3. Chong-White, N.R., Cox, R.V.: An intelligibility enhancement for the mixed excitation linear prediction speech coder. IEEE Signal Processing Letters 10(9), 263–266 (2003)
4. Martin, R., Malah, D., Cox, R.V., Accardi, A.J.: A noise reduction preprocessor for mobile voice communication. EURASIP Journal on Applied Signal Processing, 1046–1058 (2004)
5. Demiroglu, C., Anderson, D.V.: A soft decision MMSE amplitude estimator as a noise preprocessor to speech coders using a glottal sensor. In: Proc. ICSLP, pp. 857–860 (2004)
6. Quatieri, T.F., Brady, K., Messing, D., Campbell, J.P., Campbell, W.M., Brandstein, M.S., Clifford, C.J., Tardelli, J.D., Gatewood, P.D.: Exploiting nonacoustic sensors for speech encoding. IEEE Trans. on ASLP 14(2), 533–544 (2006)
7. Hu, Y., Loizou, P.C.: A comparative intelligibility study of speech enhancement algorithms. ICASSP 4(4), 561–564 (2007)
8. Liu, W.M.: Objective assessment of comparative intelligibility. PhD Thesis, University of Wales Swansea University (2008)
9. Supplee, L.N., Cohn, R.P., Collura, J.S., McCree, A.V.: MELP: The new federal standard at 2400 bps. In: Proc. ICASSP, vol. 2, pp. 1591–1594 (1997)
10. Hory, C., Martin, N.: Spectrogram segmentation by means of statistical features for non-stationary signal interpretation. IEEE Trans. on Signal Processing 50, 2915–2925 (2002)
11. Cox, R.V.: Three new speech coders from the ITU cover a range of applications. IEEE Communications Magazine, 40–47 (1997)
12. Gibson, J.D.: Adaptive prediction in speech differential encoding system. Proc. IEEE 68, 488–525 (1980)
13. Ekudden, E., Hagen, R., Johansson, I., Svedberg, J.: The adaptive multi-rate speech coder. In: Proc. IEEE Workshop on Speech Coding, pp. 117–119 (1999)
14. Chen, J.-H., Cox, R.V., Lin, Y.-C., Jayant, N., Melchner, M.J.: A low-delay CELP coder for the CCITT 16 kb/s speech coding standard. IEEE Selected Areas in Communications 10(5), 830–849 (1992)
15. Vary, P., Hellwig, K., Hofmann, R., Sluyter, R.J., Galand, C., Rosso, M.: Speech codec for the european mobile radio system. In: Proc. ICASSP, pp. 227–230 (1988)
16. Tremain, T.E.: The government standard linear predictive coding algorithm: LPC-10. In: Speech Technology, pp. 40–49 (1982)
17. Sun Microsystems. CCITT ADPCM encoder G.711, G.721, G.723, encode (14/04/2008), `ftp://ftp.cwi.nl/pub/audio/ccitt-adpcm.tar.gz`

18. 3GPP. European digital cellular telecommunication system 4750.. 12200 bits/s speech CODEC for adaptive multi-rate speech traffic channels, encoder, v6.0.0 (29/06/2008), http://www.3gpp.org/ftp/Specs/html-info/26073.htm
19. Zatsman, A., Concannon, M.: 16 kb/s low-delay CELP algorithm, ccelp, v2.0 (14/04/2008), ftp://svr-ftp.eng.cam.ac.uk/comp.speech/coding/ldcelp-2.0.tar.gz
20. Jutta. ETSI 06.10 GSM-FR, toast, v1.8 (14/04/2008), http://kbs.cs.tu-berlin.de/~jutta/toast.html
21. Texas Instruments, Inc. 2.4 kb/s proposed federal standard MELP speech coder, melp, v1.2 (14/04/2008)
22. Fingerhut, A.: U.S. department of defence LPC-10 2400bps voice coder, nuke, v1.5 (14/04/2008), http://www.arl.wustl.edu/~jaf/lpc/
23. Liu, W.M., Jellyman, K.A., Mason, J.S., Evans, N.W.D.: Assessment of objective quality measures for speech intelligibility estimation. In: Proc. ICASSP (2006)
24. ITU recommendation P.56. Objective measurement of active speech level. ITU (1993)
25. Hirsch, H.G., Pearce, D.: The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the next Millenium (2000)
26. Liu, W.M., Jellyman, K.A., Evans, N.W.D., Mason, J.S.D.: Assessment of objective quality measures for speech intelligibility. Publication in ICSLP (accepted, 2008)
27. Romero Rodriguez, F., Liu, W.M., Evans, N.W.D., Mason, J.S.D.: Morphological filtering of speech spectrograms in the context of additive noise. In: Proc. Eurospeech (2003)
28. Evans, N.W.D.: Spectral subtraction for speech enhancement and automatic speech recognition. PhD Thesis, University of Wales Swansea (2003)
29. McAulay, R.J., Quatieri, T.F.: Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans. ASSP 34(4), 744–754 (1986)
30. ImageMagick Studio LLC. Imagemagick, v6.3.0, http://www.imagemagick.org
31. Kuhn, M.: JBIG-KIT package, v1.6, http://www.cl.cam.ac.uk/~mgk25/jbigkit/