CLEAR Project

Speech Cleaning Literature Review

Mike Brookes, Nick Gaubitch, Mark Huckvale, Patrick Naylor

15/05/2008

PUBLIC RELEASE

Executive Summary

The aim of speech cleaning is to improve the intelligibility and/or the listenability of poor quality speech signals. The causes of poor quality may include additive acoustic background noise, convolutive channel effects and distortions introduced in the electronic recording chain such as clipping, electrical noise and codec distortions. This report provides a summary of the literature on three aspects of speeh cleaning. First it discusses the development and current state of methods used to evaluate the quality and intelligibility of speech. Second, it reviews the principal approaches that have been proposed for cleaning or enhancing speech signals and, third, it reviews methods for estimating the characteristics of the background noise. The report also includes a brief overview of commercial speech cleaning systems and of sound databases that may be useful in evaluating and developing speech cleaning systems.

Contents

List of Abbreviations

AI Articulation index

ALCons Articulation Loss of Consonants

AMDF Average Magnitude Difference Function. A function with similar properties to the cross- or auto-correlation but that requires no multiplications to evaluate.

AR Autoregressive

CASA Computational Auditory Scene Analysis

CCI Call Clarity Index

CCR Comparison Category Rating

CODEC Coder-Decoder

CVC Consonant-Vowel-Consonant

DAM Diagnostic Acceptabiliity Measure

DCR Degredation Category Rating

DFT Discrete Fourier Transform

DRT Diagnostic Rhyme Test

EM Estimation-Maximization. An iterative technique to solve certain optimization problems.

ETSI European Telecommunications Standards Institute

function

GMM Gaussian Mixture Model. An approximation to an arbitrary probability density function that consists of the weighted sum of several Gaussian distributions.

GSM Global System For Mobile Communications

HINT Hearing-in-Noise Test

HMM Hidden Markov Model

IRS Inverse repeated Sequence. A pseudo random sequence used for impulse response measurement.

ITU International Telecommunication Union

KLT Karhunen Lòeve Transform

LF The Liljencrants-Fant model of the gloattal flow waveform.

LLR Log-Likelihodd Ratio

LMS Least Mean Square adaptive filter

LPC Linear Predictive Coding. An autoregressive model of speech production.

LSP Line Spectrum Pairs

MARS Multivariate Adaptive Regression Splines

Mel-scale A non-uniform frequency scale that corresponds to preceived frequency. It is approximately linear at low frequencies and logarithmic at high frequencies.

MFCC Mel-frequency Cepstral Coefficients. A transformation of a signal spectrum that is obtained by taking the DCT of the log of the spectrum expressed on the mel scale.

ML Maximum Likelihood

MLS	Maximum Length Sequence of pseudo random bits.
MMSE	Minimum Mean Squared Error
MOS	Mean Opinion Score
MRT	Modified Rhyme Test
MTF	Modulation Transfer Function
NIST	National Institute Of Standards And Technology
NLMS	Normalized Least Mean Square adaptive filter
PESQ	Perceptual Evaluation Of Speech Quality
RIR	Room Impulse Response
RLS	Recursive Least Squares adaptive filter
SII	Speech Intelligibility Index
SNR	Signal-To-Noise Ratio
SPIN	Speech Perception In Noise
SPQA	Speech Quality Assurance Package
SRR	Signal-to-Reverberant ratio
SRT	Speech Reception Threshold
STFT	Short Time Fourier Transform
STI	Speech Transmission Index
STMI	Spectro-Temporal Modulation Index
STSA	Short Time Spectral Analysis
VAD	Voice Activity Detection
VCV	Vowel-Consonant-Vowel

1 Introduction

The terms “speech enhancement” and “speech cleaning” propely refer respectively to the improvement of the quality or intelligibility of speech and the reversal of degradations that have corrupted it; in practice however, most authors use the two terms interchanably. This report is concerned with cleaning methods that use only a single microphone signal; other, often more effective, techniques are possible if multiple microphone signals are available. The principal degredations that we are concerned with are (a) additive acoustic noise, (b) acoustic reverberation, (c) convolutive channel effects resulting in an uneven or bandlimited response, (d) non-linear distortion such as arises from clipping, (e) additive broadband electronic noise (f) electrical interference (g) codec distortion, (h) distortion introduced by recording apparatus.

The aims of speech cleaning vary according to the application and may include:

1. Improvements in the intelligibility of speech to human listeners.
2. Improvement in the quality of speech that make it more acceptable to human listeners.
3. Modifications to the speech that lead to improved performance of automatic speech or speaker recognition systems.
4. Modifications to the speech so that it may be encoded more effectively for storage or transmission.

Although we are, in this project, primarly concerned with the first two of these aims, most of the literature on the subject has concentrated on the last three. The distinction is important because it has been found that some approaches that are effective in improving the apparent quality may actually damage intelligibility. Despite its apparent similarity to that of improving intelligibility, improving the performance of speech recognition systems is significantly different because such systems ignore many of the cues used by humans and, in particular, normally use a coarse spectral representation of the signal from which most voicing information has been removed.

This report is a literature review covering three main areas: evaluation methods, speech enhancement and noise estimation. Section 2 reviews methods that have been developed for evaluating speech intelligibility and speech quality. For both of these tasks, it discusses both “subjective” methods that require the judgement of human listeners and “objective” methods that attempt to predict these judgements using signal analysis. Section 3 gives an overview of techniques that have been developed for single-microphone speech enhancement. An important step in several methods of speech enhancement is the estimation of the noise characteristics and, in particular, its mean power spectrum. Section 4 discusses the ways that have been used to etimate the noise characteristics from the noisy speech signal. Finally, sections **??** and 5 give a brief overview of some of the Commercial Speech Enhancement Systems and Databases that are available.

2 Evaluation Methods for Speech Enhancement Systems

2.1 Objectives

The evaluation of communication channels for speech has a long history, going back at least as far as Fletcher & Steinberg’s work on articulation testing [Fletcher and Steinberg, 1929]. The aim of this section, which draws on the excellent review in Chapters 10 and 11 of Loizou [2007], is to look at what elements of that history have applicability today for the evaluation of speech enhancement systems.

Methods for evaluation can be divided into those that look at the effect of the processing system on *Speech Intelligibility*, and those that look at its effect on the acceptability of the speech signal, known as *Speech Quality*. Methods can also be described under the heading of *Subjective* when they require the judgements of human listeners, against *Objective* when those judgements are predicted from some analysis of the signal. Subjective methods can be divided further into *Absolute* scoring methods – where a single stimulus is rated, against *Preference* methods – where multiple signals are compared. Objective methods can further be sub-divided into *Intrusive* methods – which require access to both original and processed signals, and *Non-intrusive* methods – which only require access to the processed form.

2.2 Evaluation of Speech Intelligibility

When assessing the intelligibility of a speech signal, we need to choose an appropriate linguistic level at which to make measurements. Do we want to measure the accuracy with which each phonetic element is communicated? Or whether each word is identifiable? Or whether the meaning of a sentence can be understood? The problem is one of increasing redundancy: not all phonetic sequences make up words in the language, not all word sequences make up meaningful sentences. You don't need to correctly identify all segments to identify a word, and you don't need to identify all words to understand a sentence. This linguistic redundancy introduces another problem in that individual human listeners will vary in their ability to make use of these linguistic constraints. While we may want to assess the utility of a channel to convey the meanings of real spoken utterances, listeners will vary widely in their ability to understand the speech depending on their own linguistic competence.

Thus most speech intelligibility tests are either of phonetic units made up into nonsense syllables, or of words in isolation or in short sentences.

2.2.1 Subjective Methods

2.2.1.1 Phonetic tests The original articulation test material in Fletcher and Steinberg [1929] consisted of a list of 66 CVC nonsense syllables . Since the list was balanced for each consonant and vowel phoneme, the number of syllables heard correctly gave an indication of phoneme intelligibility.

Miller and Nicely [1955] used a similar approach, but reduced the consonants to a size 16 subset: /p, t, k, b, d, g, f, θ, s, ʃ, v, ð, z, ʒ, m, n/ in a VCV frame. This subset was designed so that intelligibility could be broken down into five articulatory dimensions: voicing, nasality, frication, sibilance and place. Since then VCV materials have been widely used within speech perception research, because intelligibility results can be analysed in terms of the effectiveness of the primitive acoustic patterns necessary for consonant discrimination. However, for overall assessment of intelligibility, most effort has gone into word-level tests.

2.2.1.2 Word intelligibility One problem with the use of nonsense syllables is that listeners require training to be able to identify component phonetic units, and may be confused by phonemes which don't readily accord with spelling. Limiting listener responses to real words allows them to respond in ordinary spelling, but introduces other difficulties: firstly that different listeners may have different degrees of familiarity with the words being used and their potentially confusing competitors; secondly that words are memorable, and that, having heard a word once, listeners may be biased in their usage of the word another time. One solution to these problems include creating multiple lists of balanced difficulty, so that a listener can be used more than once. Another is to create tests with closed response sets, so every listener needs to make the same choices about the word under test.

One of the earliest word lists created for intelligibility testing was that of Egan [1948], who constructed 20 lists of 50 monosyllabic words. His aim was to balance average difficulty and range of difficulty over the lists, while ensuring that the phonetic units present were equally represented. Egan introduced the idea of "phonetic balance" by which he meant that the relative frequency of phonemes in the word lists matched the relative frequency of phonemes in conversational speech. Lehiste and Peterson [1959] built a similar set of lists of monosyllabic words, but which were balanced for the relative frequency of phonemes in monosyllabic words. Tillman et al. [1966] built on this idea to construct the Northwestern University Auditory test 4 which was used for audiometry. Similar word lists were constructed in British English by Fry and Kerridge [1939], Dickson and Chadwick [1950] and Fry [1961].

Fairbanks [1958] modified Egan's idea so that instead of identifying the whole word, listeners merely had to identify the leading consonant (the rest of the word was written down on the response sheet), leading to the concept of a "rhyme" test. House et al. [1965] modified Fairbanks' idea by introducing a closed response set, where listeners only had to choose amongst six alternatives: this was called the Modified Rhyme Test (MRT). Kreul et al. [1968] attempted to modify the MRT to make it a clinically-useful tool, while a phonetically-balanced version of the MRT was developed by Clarke [1965]. The MRT procedure has also been used for intelligibility testing in British English [Haggard and Mattingly, 1968].

In the Diagnostic Rhyme Test (DRT), Voiers [1983] extended the MRT by requiring that listeners choose between a pair of words differing in only one phonetic feature. Voiers chose features that had appeared in the analysis by Miller and Nicely [1955] of consonant confusions. This allowed listener judgements to be analysed in terms of their performance on the different feature dimensions as well as on their overall score. Extensive testing of the DRT has shown that it has high reliability, with repeatability being within 1% for a group of 10 listeners [Voiers, 1983].

2.2.1.3 Sentence intelligibility Words in isolation are not typical of the way in which speech is used for communication. In everyday listening to speech, knowledge of the topic, semantic, grammatical and prosodic constraints provide additional information which allow us to recognise words. However the use of sentences in intelligibility testing only amplifies the problems discussed earlier in relation to the use of words: listeners will vary even more in their ability to extract higher level information from sentences to aid recognition, and sentences are even more memorable. It is also much harder to create balanced lists of sentences, since the amount of benefit to be gained from sentence context is hard to quantify.

Kalikow et al. [1977] developed the Speech Perception in Noise (SPIN) test to try and assess the extent to which listeners were able to extract contextual information from sentences to aid word recognition. The SPIN test comprises 8 lists of 50 sentences, with one key-word per sentence. There are 25 key words in each list, occurring once in a high predictable context (e.g. “The boat sailed across the bay”), and once in a low predictable context (e.g. “John was talking about the bay”). The relative scores on the two types of sentence provide a measure of the listener’s ability to use context. A number of studies [Bilger et al., 1984, Morgan et al., 1981] have suggested improvements in the design of the SPIN test.

The Hearing in Noise test (HINT) is a revised version of the SPIN test developed by Nilsson et al. [1994]. The HINT test comprises 25 balanced lists with 10 sentences per list. All words in the list are scored, but alternatives are allowed for grammatical words (e.g. a, the, is, was, . . .). A standard recording of the lists by a professional male actor has been made available and the HINT test remains popular for the assessment of hearing impairment.

Two testing methodologies are in common use for the measurement of intelligibility with speech materials such as HINT. In the first method, the speech signals are first processed into a number of fixed levels of presentation: where the level is either the speech amplitude or the signal to noise ratio. Then speech materials from all levels are presented to the listeners in random order, and the intelligibility at each level is recorded and plotted as a function of the presentation level. This graph usually takes the shape of an S-shaped psychometric function as intelligibility changes from 0 to 100% across a range of levels. From this graph we can estimate the presentation level which achieves a required level of intelligibility. The level at which 50% intelligibility is obtained is called the Speech Reception Threshold (SRT) [Plomp and Mimpen, 1979].

This fixed presentation method has a number of disadvantages: firstly, the same amount of speech is presented at every level - even though we may be only interested in one point on the psychometric curve – this makes the test rather long; secondly, to ensure that the levels bracket 50% intelligibility say, we may need to use a wide range of levels, many of which will have 0% or 100% intelligibility for a particular listener; thirdly, we need to ensure that the speech material presented at each level is equally difficult.

An alternative to this method is to use an adaptive procedure where presentation levels change dynamically within the test to determine the level which delivers an intelligibility score of required size. In a typical procedure, the speech material is first played at a good level/SNR such that listeners can readily identify the words. Then the level is adjusted such that the presentation level/SNR falls by (say) 2dB when the listener gets the words correct, and rises by 2dB when the words are incorrect. Typically the average of the levels at which the last eight reversals in direction were made is taken to be the overall SRT score. Studies into the effectiveness of this procedure has shown that it seems to work well [Levitt, 1971, Dirks et al., 1982].

A disadvantage of an adaptive procedure is that it relies on assumptions about the shape of the psychometric function. A curve that is not very steep at 50% intelligibility will produce more variable results, and a curve that is not monotonic may not produce sensible results at all.

Where the systems under test are available as real-time implementations, it is possible to process stimuli while a test is taking place; this avoids the need to capture and process the stimuli off-line [Worrall et al., 2006]. In combination with an adaptive procedure, this makes intelligibility testing fast enough for rapid assessment of systems in the field.

2.2.2 Objective Methods

Objective measures of the intelligibility of a channel are usually based on the premise it is the addition of reverberation and noise that has the most significant impact on intelligibility. Through comparison of the original and processed signals and working from the assumption that the original is 100% intelligible, these methods attempt to estimate the loss in intelligibility caused by the addition of reverberation or noise. Over the years several objective measures of intelligibility have been developed. Most are based on measurements of how changes to the signal are distributed across a number of spectral bands. This idea was first proposed in French and Steinberg [1947] and led to the definition of the Articulation Index (AI) [Kryter, 1962, ANSI, 1969] and the Speech Intelligibility Index (SII) [ANSI, 1997]. In this approach measurements of the speech and noise levels in different spectral bands are weighted according to the relative contribution of the band to intelligibility. Evaluation has shown that such measurements correlate well with subjective intelligibility scores, at least for steady-state noise and normally-hearing listeners. Recent extensions to the SII approach have attempted to account for the loss in intelligibility caused by fluctuating noise [Ludvigsen, 1985, Rhebergen and Versfeld, 2005, Rhebergen et al., 2006].

The Speech Transmission Index (STI) is a formalisation of the SII developed by Steeneken and Houtgast [Steeneken and Houtgast, 1980, Houtgast and Steeneken, 1985, Steeneken and Houtgast, 1999]. Instead of using SNR to measure the audibility of speech in each spectral band, the STI uses a modulation transfer function (MTF) to assess the degree to which intensity modulation in each channel is preserved. STI measurements are usually made by passing a special speech-like signal through the channel under test. Reductions in the modulation depth of the test signal at the output of the analysis channels indicates a loss in information. Evaluations have shown that the STI makes good predictions for the intelligibility of speech disturbed by linear filtering, reverberation or the addition of steady-state noise [Houtgast and Steeneken, 1985]. However the measure is not suitable for non-linear processes, phase jitter, compression, peak limiting or centre-clipping.

To model the effect of non-linear processes on intelligibility, Elhilali et al. [2003] developed a modified MTF called the spectro-temporal modulation index (STMI). As well as taking into account changes in temporal modulations of intensity within a channel, like the MTF, the STMI also measures spectral modulations across channels.

Other methods, not based on the articulation index approach, have also been used for intelligibility testing. The %ALCons measurement is a widely used predictor of intelligibility in buildings that takes into account the effect of reverberation and distance to source. The model of Holube and Kollmeier [1996] predicts the intelligibility of CVC words through comparison with a database of word templates. Words become less intelligible when the processing makes them more similar to competitors in the database. Finally, intelligibility can be predicted from a database of real intelligibility measurements by comparing actual channel conditions with conditions used during testing. For example, Fletcher and Galt [1950] built a mathematical predictor from intelligibility data collected across many listening conditions.

Objective speech quality measures, such as PESQ, have also been used as predictors of intelligibility and an evaluation of a number of measures can be found in Liu et al. [2006] . Yamada et al. [2006] built a predictor for intelligibility from PESQ predictions of MOS for a number of noise conditions processed by a number of noise-reduction algorithms. The effectiveness of prediction varied with word familiarity, but correlations with subjective intelligibility scores as high as 0.9 were obtained.

Speech recognition systems have also been used to obtain objective predictions of speech intelligibility, for example Chernick et al. [1999] contrasted the performance of a phone recogniser on the original and degraded signals, to establish a prediction for the loss in intelligibility. Liu et al. [2006] compared a speech recognition approach with other objective measures and found that it gave good intelligibility prediction at +10dB SNR, but was significantly worse than other measures at lower values, probably because human recognition performance is substantially better than a recogniser at low SNR.

2.3 Evaluation of Speech Quality

The assessment of a communication channel for its effect on speech intelligibility is only one part of the problem of system evaluation. Two channels which provide the same subjective intelligibility may nevertheless differ in other ways, and in particular may be more or less acceptable to listeners. This assessment of acceptability of a channel, has become known as the evaluation of speech quality, although the term is rather a confusing one, since often it is the degree to which background and system noise interferes with the speech that is being assessed.

2.3.1 Subjective Methods

The assessment of speech quality is more difficult than that of speech intelligibility, since the reliability of listeners becomes an issue. There are broadly two types of speech quality test: one in which listeners assign absolute ratings to individual speech stimuli, and one where listeners exhibit a preference for one speech stimulus over one or more others. The advantage of the first type is that a system can be assigned some absolute score, a disadvantage is that large numbers of listeners are required to achieve satisfactory sensitivity. The advantage of the second type is that statistically significant results can be obtained from a comparison of two systems with relatively few listeners [Vazquez-Alvarez and Huckvale, 2002], a disadvantage is that an estimate of the size of the difference between the systems may not be available.

2.3.1.1 Quality rating tests A widely used rating scale for the assessment of speech quality is shown in Table 1. Standard procedures have been published about how the scale should be used within

Rating	Speech Quality	Level of Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible, but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying , but not objectionable
1	Bad	Very annoying and objectionable

Table 1: Absolute Category Rating scale for Listening Quality (ACR LQ) [ITU, 1998a]

a testing procedure [Rothauser et al., 1969, ITU, 1990, 1998a]. In a training phase, listeners are played a set of reference signals demonstrating the different quality levels. This "anchoring" phase is important to standardise responses across listeners. Then the test signals are played and rated by listeners. Recommendations are that at least 20 listeners are used, that sessions should not last more than 20 minutes, and that stimuli be played over headphones. Raw scores can be standardised by asking listeners to rate a set of standard SNR-varying stimuli. This allows the mean opinion score (MOS) to be converted to a SNR equivalent number in dB.

A disadvantage of the MOS approach is that it does not provide any diagnostic information about which aspects of the signal were most important to listeners. Scientific studies of our perception of distortion [McDermott, 1968] appear to show that the quality of distorted speech is perceived along dimensions of overall clarity, signal-background distortion, and subjective loudness. The Diagnostic Acceptability Measure (DAM) developed in Voiers [1977] builds on this analysis to request that listeners rate the quality of speech stimuli along a number of dimensions independently, ignoring other aspects. For example, listeners may be asked to rate the speech solely as "nasal", or as "muffled"; or they may be asked to rate the background as "buzzing" or "rumbling". Although such a test provides more information about the criteria that listeners use to rate the quality of a channel, a disadvantage is that the testing is time-consuming and that listeners need to be trained and anchored in each dimension.

The assessment of speech that has been passed through noise-suppression or signal enhancement algorithms creates an additional complexity, since it is not clear whether the overall rating of the channel by listeners is because of the effect of the processing systems on the speech, on the noise or on both. This has lead to a recommendation for testing signal enhancement systems [ITU, 2003], whereby individual assessments of the speech signal and the background are requested of listeners in addition to an overall rating. Each speech stimulus is played three times: first listeners are asked to judge the speech itself on a 5-point scale of signal distortion; then listeners are asked to judge the background on a 5-point scale of background intrusiveness; finally listeners are asked to judge the overall effect using the 5-point MOS scale. The types of impairment that arise in telephone systems and their influence on MOS scores is discussed in ITU-T [1993b].

2.3.1.2 Preference tests In preference testing, listeners are typically presented with pairs of speech stimuli, differing only in the processing system. The reference stimuli can come from a second system, in which case the test can be used to compare two systems, or they can come from a set of reference processing conditions, in which case the test system is compared to a known set of distortions. Examples of these are Munson and Karlin [1962] and Hecker and Williams [1966]. A preference system that used five distortion conditions (clean, band-pass filtered, low-pass filtered, reverberant, peak-clipped) was proposed as a standard by the IEEE subcommittee for speech quality measurement [Rothauser et al., 1969].

Another technique for generating a set of reference stimuli to use in preference testing was described in Combescure et al. [1982]. Here, reference signals were generated using modulated noise [Law and Seymour, 1962], and listeners were asked to rate their preference of test stimuli over the reference on a five-point scale, called the Degradation Category Rating (DCR) scale. The use of a rating scale when making preference judgements can also be used even when comparing two systems: with a scale from –3 to +3 for example. This is then called the Comparison Category Rating (CCR) method.

2.3.2 Objective Methods

Although subjective testing is the only way to obtain true judgements of speech quality, a number of objective methods have been developed, which despite their shortcomings, have been found to correlate well with subjective ratings, such as the MOS. For a review of objective speech quality measures, see Quackenbush et al. [1988].

2.3.2.1 Intrusive Intrusive measures of speech quality compare the processed speech signal with the original clean signal that went in to the communication channel. Typically the process involves dividing the signals into short frames of 10-30 ms, and computing a distortion measure between equivalent frames. An average is then taken over all frames to obtain a global measure. Examples are in Brandenburg [1987] and Schroeder et al. [1979].

There are many variants of the Segmental SNR approach. Here aligned frames are compared either as time-domain waveforms, or as spectra, and differences between input and output are weighted according to the true size of the input. An advantage of using a frequency domain measure is that different weights can be applied to different frequency bands.

A wide variety of spectral distance measures have been proposed to assess the significance of the distortion caused by the channel. The log-likelihood ratio (LLR), Itakura-Saito, and cepstral distance measures have all been applied to LP-modelled spectra from the input and output frames [Quackenbush et al., 1988].

In recent years, perceptually motivated measures have dominated techniques for objective measurement of speech quality. The attempt here is to model the signal processing that takes place in the peripheral auditory system, including filtering into auditory channels, compression, loudness pre-emphasis, subjective loudness compensation and aspects of masking in time and frequency. In a competition to design a new objective measure capable of performing reliably across a wide range of speech coders, the ITU chose the PESQ algorithm from Rix et al. [2001] to form the currently recommended technique ITU-T P.862 [ITU, 2001]. PESQ consists of a complex sequence of processing steps to generate a set of distortion scores as a function of frequency and time. The steps include: amplitude equalisation, telephone handset filtering and time alignment, then an auditory transformation which comprises Bark spectrum estimation, frequency equalisation, gain equalisation and loudness computation. The difference between the loudness spectra arising from the output signal are then compared to the loudness spectra arising from the original signal are used to estimate a “disturbance” score, which is then adjusted to account for masking effects, and the perceptual difference between gains and losses in energy. The disturbance scores are then averaged across frequency and time taking into account potentially invalid values caused by signal mis-alignment. The result is a PESQ score on a scale between 0.5 and 4.5 which has been shown to correlate well with subjective listening tests over a range of telephony channels [Rix et al., 2001]. PESQ has been evaluated specifically for noisy speech processed by noise-reduction algorithms by Kitawaki and Yamada [2007], where correlations between 0.83 and 0.96 were obtained across a variety of noise types and processing algorithms.

Composite methods have been used to achieve a better match to subjective scores. A number of objective methods can be combined using statistical techniques, such as multivariate adaptive regression splines (MARS) described in Friedman [1991]. A composite objective measure for the objective quality rating of speech enhancement systems is described in Hu and Loizou [2006]. Five objective measures are compared with P.835 [ITU, 2003] subjective measures for a range of noise, SNR and enhancement algorithms. Results showed that some measures, such as segmental SNR, were not good predictors of the quality of enhanced speech. Multiple linear regression was used to obtain the best predictor for each of the subjective measures.

2.3.2.2 Non-intrusive In some situations where objective quality measures are useful, the original clean signal is not available. Assessment must then be based on the processed signal alone. A review of a number of techniques for non-intrusive objective assessment of quality is given in Rix [2004].

For telephone circuits, non-intrusive measures of speech level, noise level, talker echo and delay can easily be measured and used to make some prediction of the likely channel quality, for example the Call Clarity Index (CCI) described in ITU-T P.562 [ITU, 2004a].

To obtain single-ended measures of speech quality, it is necessary to analyse the degree to which the signal appears to follow the typical statistics of speech. For example Gray et al. [2000] analysed the signal as a sequence of predicted vocal tract shapes, then rated the plausibility of the shapes and transitions. Another approach, that works from the auditory processing stages of the PESQ approach, was described in Beerends et al. [2000]. A recent competition for non-intrusive quality models as organised by ITU-T led to recommendation P.563 for a perceptual single-sided speech quality measure[ITU, 2004b]. More recently a number of algorithms have been developed that claim to give significantly better performance than P.563. Grancharov et al. [2006] developed a low-complexity measure that gave superior prediction of MOS scores with much less computational cost than P.563, while Kim and Tarraf [2006] describe the ANIQUE+ model which is trained on the MOS results found for 24 different speech databases covering a wide variety of distortion conditions. They claim their model predicts MOS performance better even

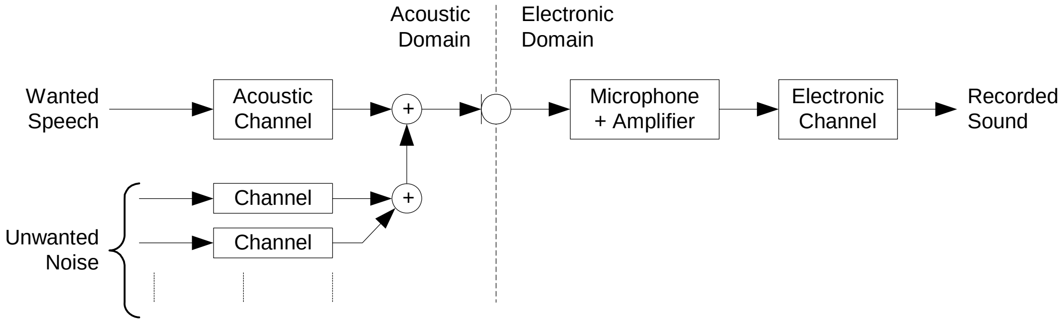

Figure 1: Typical speech recording chain.

than the P.862 intrusive method.

2.4 Issues specific to the assessment of enhancement systems

Most evaluation measures have been designed to assess the performance of telecommunication channels, and it could be argued that the assessment of enhancement system creates new challenges. One issue, for example, arises from the non-linear processing that takes place within enhancement systems so that the types of distortion that occur when noise is suppressed may not be typical of conventional channels. Another factor is that most enhancement systems adapt their processing to signal properties, or they may have many parameters that need to be set: thus it is hard to ensure that a processed stimulus is a fair representation of the capabilities of the system.

Hansen and Pellom [1998] propose a standard protocol for the assessment of enhancement systems which uses (i) a standard set of sentences, (ii) a standard set of noise types, (iii) a combination of objective and subjective assessment methods. The designer of a speech enhancement system can use this protocol to provide a reference against which to assess his own system.

Another problem might arise with the use of the SRT measurement framework, where the adaptive procedure is used to adjust the SNR of the speech going in to the enhancement system. The aim may be to find the input SNR which achieves 50% intelligibility at output. However, since the systems are adaptive, and almost certainly make assumptions about the input signal, the shape of the performance-SNR curve may not be a typical psychometric function, indeed, many enhancement systems do not perform well on clean speech and this performance curve may not even be monotonic with SNR.

3 Speech Enhancement

Signal degradations may arise at any point in the recording chain illustrated in Fig.1. It is convenient to categorise these degradations into the three groups given below according to the way in which they alter the wanted speech signal; this categorisation also corresponds approximately to their perceived effect and to the appropriate cleaning method.

1. Additive noise that is uncorrelated with the wanted speech signal may arise in either the acoustic or electronic domain. Its perceived effect is to degrade listenability and intelligibility and may, in extreme cases, completely mask the wanted signal. For some types of additive noise, the spectral characteristics are stationary or change slowly with time. This is typically true of hum and amplifier noise as well as of some environmental acoustic noise sources. Spectral subtraction (Sec. 3.2) and single-channel adaptive filtering (Sec. 3.5) have been successful in reducing the perceived level of such stationary noise sources. Other forms of additive noise are intermittent or highly non-stationary and their identification and deletion is the subject of model-based and missing data methods (Sec. 3.6 and 3.8). Such non-stationary noise sources include media interference, unwanted co-talkers and some forms of electrical interference.

2. Convolutive effects are perceived as reverberation and poor spectral balance; they differ from the previous group because the added noise is strongly correlated with the wanted signal. Reverberation and echo normally arise from acoustic reflections and can seriously degrade intelligibility. The increasing use of distant microphones in hands-free telephony has prompted extensive research into

reducing the effects of reverberation (Sec. 3.11). Bandwidth restrictions and uneven spectral response may arise from microphone placement, microphone characteristics and CODEC limitations. There has been some work on expanding the bandwidth narrow-band telephone signals (Sec. 3.10) in order to improve listenability but there is little evidence of any intelligibility benefits.

3. Non-linear distortion frequently arises from amplitude limiting or clipping in the microphone, amplifier or CODEC. This is perceived as harsh distortion that varies with the signal amplitude. A similar perceptual effect can result from high bit errors in the coded signal used by some CODECs. Clipped portions of a waveform are easy to identify provided that no subsequent phase distortion is present and techniques exist for reconstruction the corrupted portions of the waveform (Sec. 3.9).

Most speech enhancement techniques address only one of the degradation mechanisms listed above although some model-based methods (Sec. 3.6) address the first two jointly.

A number of books devoted to speech enhancement have been published including Davis [2002], Benesty et al. [2005] and Loizou [2007]. In addition, there are a number of survey papers on the subject including Lim [1986], Ephraim and Cohen [2006] and Ephraim et al. [2005]. The performance of several enhancement algorithms was tested by Song et al. [2006] by means of a standardised speech recogniser; they found that some enhancement techniques reduced the recognition performance. A subjective assessment of 13 enhancement algorithms is given in Hu and Loizou [2007] which, following ITU [2003], asked listeners for separate judgements on signal distortion, backgound noise and overall quality. The model-based algorithms (Sec. 3.6) generally gave the best overall quality although a multi-band spectral subtraction algorithm (Sec. 3.2) developed by the authors themselves also performed well.

3.1 Frame Based Processing

Enhancement methods that operate in the frequency domain normally require the sampled input signal $x(n)$ to be decomposed into overlapping frames with the ℓth frame given by

$$x(n;\ell) = w(n)x(n + \ell M), \tag{1}$$

for $n = 0, \cdots, N-1$ where $w(n)$ is a windowing function with finite support and $M \le N$ is the time increment between successive frames (in samples). The window length, N, is a compromise between frequency and time resolution and is typically chosen in the range $10-30$ ms for speech signals resulting in a frequency resolution of around 50 Hz. The frame increment, M, is most commonly set at $N/2$ although, as noted below, there are theoretical reasons for using $M = N/4$ despite its higher computational cost.

A Fourier transform is normally performed on each frame to obtain the Short Time Fourier Transform (STFT). If no processing is done on the frame spectra, the original time domain signal can be reconstructed exactly with overlap-addition [Allen and Radiner, 1977, Allen, 1977]. However, when frequency-domain processing is performed on the frames, distortion artifacts may be introduced due to signal discontinuities at frame boundaries and aliasing of rapidly changing spectral coefficients. The reconstruction properties can be controlled by the choice of the windowing function and the ration $M : N$. Martin and Cox [1999] suggest the use of half-overlapping ($M = N/2$) square-root Hanning windows for both analysis and synthesis in order to provide perfect reconstruction in the absence of any processing and, at the same time, to attenuate any frame-boundary discontinuities. An extensive discussion of these issues is given in Allen [1977] and Allen and Radiner [1977] where it is shown that, for a Hamming analysis window, a three-quarters overlap ($M = N/4$) is needed in order to ensure that the spectral coefficients are sampled frequently enough to avoid aliasing.

3.2 Spectral Subtraction

The spectral subtraction method of speech enhancement was introduced in Boll [1979] and remains one of the most widely used ways of reducing additive noise; a good overview of the method and its variants is included in Virag [1999]. In the simplest form of spectral subtraction, the estimated magnitude spectrum of the noise, $|\hat{\mathcal{N}}(e^{j\omega})|$, is subtracted from that of the noisy speech to obtain the estimated magnitude spectrum of the clean speech while the phase of each spectral component is left unaltered. This process is illustrated in Fig. 3.2 and can be written in the frequency domain as

$$\hat{S}(e^{j\omega}) = G(e^{j\omega})X(e^{j\omega}) \tag{2}$$

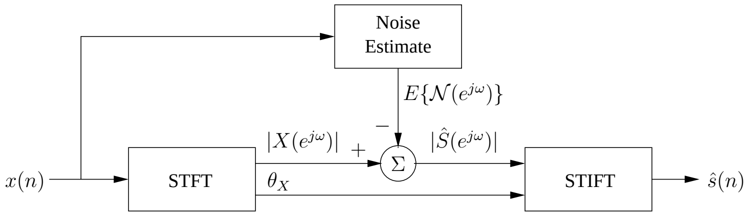

Figure 2: Spectral subtration with $\gamma = 1$.

where the real-valued gain function, $G(e^{j\omega})$, is given by

$$G(e^{j\omega}) = \max\left\{\frac{|X(e^{j\omega})| - |\hat{\mathcal{N}}(e^{j\omega})|}{|X(e^{j\omega})|}, 0\right\} \tag{3}$$

in which the max() function prevents G from becoming negative at low SNRs. Because the gain function is real-valued, the phase spectrum is uncorrected, however Wang and Lim [1982] demonstrated that little perceptual improvement resulted from using the true phase spectrum of the clean speech signal. Methods of estimating the noise spectrum, $\hat{N}$, are discussed in Sec. 4.

Berouti et al. [1979] used a more general expression for the gain function

$$G(e^{j\omega}) = \max\left\{\frac{\left(|X(e^{j\omega})|^\gamma - |\hat{\mathcal{N}}(e^{j\omega})|^\gamma\right)^{1/\gamma}}{|X(e^{j\omega})|}, 0\right\} \tag{4}$$

and investigated the cases of $\gamma = \{0.5,\ 1,\ 2\}$. Despite their finding that the best results are obtained with $\gamma = 2$, the most popular choice remains $\gamma = 1$.

Subtracting the expected noise spectrum rather than its instantaneous value causes two problems: (i) there is residual broad-band noise after processing and (ii) individual narrow band spectral spikes remain and generate tonal noise often referred to as musical noise. A number of improvements have been proposed to circumvent these problems including the introduction of a gain floor and oversubtraction [Berouti et al., 1979, Lim and Oppenheim, 1979] which result in the modified gain function,

$$G(e^{j\omega}) = \max\left\{\frac{\left(|X(e^{j\omega})|^\gamma - \alpha|\hat{\mathcal{N}}(e^{j\omega})|^\gamma\right)^{1/\gamma}}{|X(e^{j\omega})|}, \beta|\hat{\mathcal{N}}(e^{j\omega})|\right\} \tag{5}$$

where $\alpha \geq 1$ and $0 \leq \beta \ll 1$ are coefficients controlling the oversubtraction and the noise floor respectively. By setting the noise floor coefficient to $\beta > 0$ some broadband noise remains which reduces the perception of musical noise [Berouti et al., 1979, Virag, 1999]. Berouti et al. [1979] set β to a constant in the range $0.005 - 0.1$. Increasing the oversubtraction coefficient, α, reduces the residual noise but may introduce distortion of the speech signal if set to high Under the assumption that the noise power is constant while the speech power varies from frame to frame, the oversubtraction coefficient α may be varied in each frame based on the SNR in each frame so that less subtraction is performed in frames with high SNR.

Lockwood and Boudy [1992] use a frequency-dependent oversubtraction coefficient which is adaptively updated using a nonlinear function of the smoothed SNR estimate and the peak noise level in recent frames. This approach is extended in Virag [1999] where both the oversubtraction and the noise floor coefficients are controlled adaptively. Instead of SNR extrema, a perceptual threshold function is used derived from the Bark spectrum of a roughly enhanced speech signal using standard spectral subtraction.

McAulay and Malpass [1980] observe that spectral subtraction performs poorly when there is no speech present and introduce a two-state model for speech presence. Using a Gaussian noise model for the noise, they derive an expression for the probability of speech presence based on the true (or “a priori”) SNR and the ratio of noisy-signal to noise power (the “a postiori” SNR). The gain function in (4) is then multiplied by the probability of speech presence which results in greater attenutation when speech is absent. Hansen [1991] proposes an alternative approach to eliminating musical noise by borrowing morphological operations from image processing. The idea is to apply dilation or closure (dilation followed by erosion) operators to the time-frequency “image” that results from spectral subtraction in order to smooth out the residual noise without blurring speech features. To avoid the latter, he performs segmentation into voiced/unvoiced/transition/silence and only applies the morphological operations within a segment. The paper is rather fuzzy on the details. Instead of using a DFT for spectral analysis, Lin et al. [2002b] and Lin et al. [2002a] use a filterbank that approximates the critical bands of the ear and, in calculating the Wiener filter response, use the difference between the noise (estimated using minimum statistics) and the calculated masking threshold.

3.3 Minimum mean square estimators (MMSE)

In an influential paper Ephraim and Malah [1984] proposed an optimal MMSE estimation of the short-time spectral amplitude (STSA); its structure is the same as that of spectral subtraction but, in constrast to the Wiener filtering motivation of spectral subtraction, it optimizes the estimate of the real rather than complex spectral amplitudes. Central to their procedures is the estimate of SNR in each frequency bin for which they proposed two algorithms: a maximum likelihood approach and a "decision directed" approach which they found performed better. The maximum likelihood (ML) approach estimates the SNR (or "a priori" SNR) by subtracting unity from the low-pass filtered ratio of noisy-signal to noise power (the "a postiori" or "instantaneous" SNR) and half-wave rectifying the result so that it is non-negative. The decision-directed approach forms the SNR estimate by taking a weighted average of this ML estimate and an estimate of the previous frame's SNR determined from the enhanced speech; the weights used were 0.02 and 0.98 respectively. Both algorithms assume that the mean noise power spectrum is known in advance (see Sec. 4). Cohen [2004] has proposed modifications to the decision-directed approach which are claimed to improve performance further and showed that a delayed response to speech onsets could be avoided by making the estimator non-causal. Subsequently, Ephraim and Malah [1985], introduced an improved version of their procedure which minimized the mean square error of the log spectrum, rather than that of the power spectrum itself. They reported that this gave noticibly lower background noise levels without introducing additional distortion. An analysis of why the Ephraim and Malah [1984] approach gives less musical noise than spectral subtraction is given in Cappe [1994].

Manohar and Rao [2006] use techniques taken from missing feature estimation (see Sec. 3.8) to improve the performance on non-stationary noise. If the frame SNR is less than 10 dB, they assume that only the voiced components of speech will be above the noise floor. They therefore classify broad subbands as speech or noise using one of several spectral flatness measures and apply additional attenuation in the "noise" subbands. They claim that this approach works well for noise consisting of a stationary component added to impulsive bursts; the MMSE enhancer removes the stationary component while the postprocessor identifies and attenuates the impulsive bursts.

A number of authors, including Hansen et al. [2006], have incorporated a perceptual masking model into the enhancement process in order to avoid attenuating noise components that are already inaudible due to masking.

3.3.1 Super Gaussian Estimates

If the correlation length of the speech signal is larger than the analysis window, the speech spectral amplitudes will not be Gaussian and a MMSE estimator is not optimal. Lotter et al. [2003] derived optimal amplitude estimators using a Rayleigh prior for the noise spectral amplitudes (i.e. a Gaussian prior for the complex amplitudes) and an expression for the speech prior that, with appropriate parameter settings, could approximate either a Laplacian or Gamma distribution. They demonstrate (see Fig. 3) that the true histogram of spectral amplitudes does not follow a Rayleigh distribution but instead lies between a Gamma and Laplace distribution. The found that the use of super-gaussian priors gave increased noise reduction but with a slight increase in speech distortion at SNR levels below 5 dB. A similar conclusion was reached by Chen and Loizou [2005] who derived an exact, but very complicated, expression for the optimum estimator using a Laplacian speech prior.

Martin [2005] derived optimal estimators of the complex spectral amplitudes for Gamma and Laplacian priors and found that the latter resulted in less musical noise. He too found that the use of a super-gaussian speech prior gave increased noise supression but with poorer noise quality.

3.4 Subspace Methods

Acoustic models of the vocal tract justify the widely used model of speech as arising from a low-order autoregressive (AR) process normally treated as time-invariant over intervals of around 20 ms. A consequence of this model is that the speech samples within a frame of this length lie within a low-order subspace; the aim of subspace speech enhancement methods is to identify this subspace and constrain

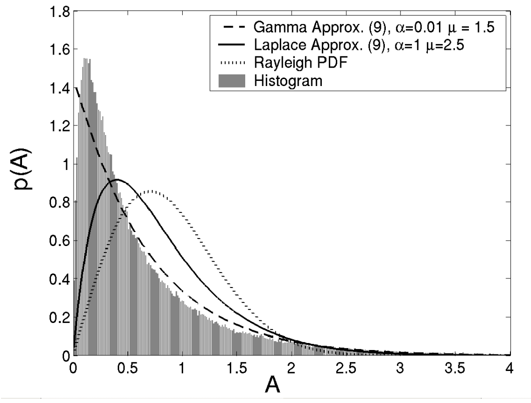

Figure 3: Hisogram of speech spectral amplitudes compared to Gamma, Laplace and Rayleigh distributions [Lotter et al., 2003].

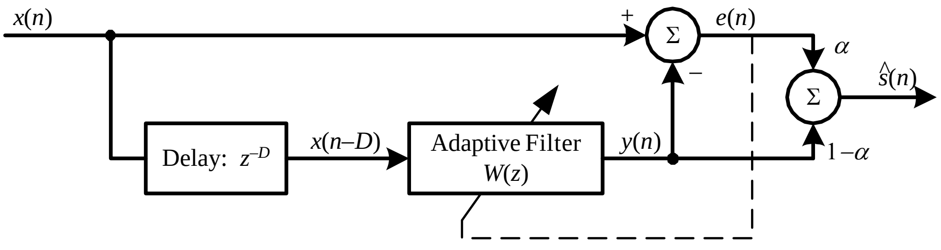

Figure 4: Adaptive noise cancellation

the clean speech samples to lie within it. The approach was introduced in Dendrinos et al. [1991] and became popular following Ephraim and Van Trees [1995] who used an eigendecomposition of the the autocovariance matrix of the input speech signal to identify the signal subspace and its complementary noise subspace. The method assumed that the noise was white and that the autocovariance matrix of the noisy speech therefore consisted of the sum of a low-rank matrix arising from the speech and a multiple of the identity matrix arising from the noise. They presented linear estimators of the clean speech which minimized the distortion of the speech subject to constraints on the noise power in either the time or frequency domain. An alternative formulation, claimed to have numerical advantages and to allow a recursive procedure, was presented in Jensen et al. [1995] and in Hansen [1997] where an efficient method of identifying the noise subspace was developed.

The original approach copes akwardly with coloured noise and a development is given in Hu and Loizou [2002b,a] where a single non-orthogonal transformation is used to diagonalize both the speech and noise covariance matrices; this approach was subsequently generalized in Lev-Ari and Ephraim [2003]. An alternative approach to dealing with coloured noise, the Rayleigh Quotient method, is given in Mittal and Phamdo [2000] and Rezayee and Gazor [2001].

The subspace enhancement algorithms make an explicit compromise between signal distortion and noise and several authors have suggested basing this compromise on perceptual models to permit higher noise in spectral regions where it will be masked, thereby allowing a reduction in distortion [Hu and Loizou, 2003, ?, Kim et al., 2003].

A more recent approach that does not rely on any explicit model of noise or speech is to assume that the speech signal lies in a low-order manifold that is embedded within a high dimension phase space (typically of dimension 20 to 30). If, as in Hegger et al. [2001] and Johnson et al. [2003], the phase space is formed from multiple delayed versions of the signal, then the approach is closely related to that of the previous paragraph.

3.5 Single-channel Adaptive Filtering

A single channel adaptive filter can be used to identify components of a signal that are correlated with previous samples; such correlations arise in particular from periodicity in the speech or noise. The general structure of a single-channel adaptive filter is shown in Fig. 4. The noisy speech signal, $x(n)$, is delayed by D samples and passed through the filter to give $y(n)$ which is subtracted from $x(n)$ to generate the error signal $e(n)$. The filter response is adjusted via the feedback path to reduce the power of $e(n)$ and the output of the filter is formed as a mixture of $e(n)$ and $y(n)$ according to whether the periodic signal components should be enhanced or supressed. For a stationary input signal $x(n)$, the impulse repsonse, $w(n)$, of the filter that minimizes the power of $e(n)$ is given by

$$\mathbf{w} = \mathbf{R}^{-1}\mathbf{g} \tag{6}$$

where $r_{i,j} = E\left(x(n)x(n+i-j)\right)$ and $g_i = r_{i,-D} = E\left(x(n)x(n+i+D)\right)$ for $i, j \geq 0$. It is possible for the delay, D, to be negative but in this case we must set $w(-D) = 0$ and must introduce a delay of D samples in the input to ensure that the system is realisable. The frequency resolution of the filter is approximately equal to the reciprocal of its impulse response length in seconds.

The filter adaptation is most often performed on the $w(n)$ directly using either the LMS or, more commonly, the NLMS gradient descent algorithm [Haykin, 2002]. The LMS and NLMS algorithms can be made "leaky" by making the coefficients decay to zero over time [Mayyas and Aboulnasr, 1997]; this may improve performance with non-stationary inputs such as speech signal. Alternatively, it is also possible to use the RLS or sliding-window RLS algorithms [Haykin, 2002] which, at the cost of additional computation, recursively solve equation (6) exactly. The computation can be reduced by processing the signal in blocks and evaluating (6) only once per block. It is also possible to implement the adaptive filter using a lattice structure and to adapt the lattice coefficients using gradient descent [Friedlander, 1982].

The use of adaptive filters for noise reduction was introduced by Widrow et al. [1975]. Although their primary concern was with two channel adaptive filtering, in which a separate noise reference signal is available, they also discuss two applications of the single channel configuration shown in Fig. 4. In the first of these, the removal of periodic noise from a broadband signal, the output mixing factor is $\alpha = 1$ and the delay D is set to be long enough so that $s(n)$ and $s(n+D)$ are uncorrelated. Their second application is the reverse of the first and aims to remove broadband noise from a periodic signal; in this case $\alpha = 0$ to give $\hat{s}(n) = y(n)$. A complication of using single channel adaptive filters for speech enhancement is that both periodic and broadband components are often present in both speech and noise; it is therefore necessary to select the parameters of the adaptive filter carefully to enhance only the wanted components.

Sambur [1978] aimed to enhance the periodic component of voiced speech ($\alpha = 0$) and used an external pitch detector to set the delay D to equal a pitch period. He also used a VAD to detect the presence of speech and inhibited adaptation when speech was absent. With white noise interference (the best case), he reported an improvement of 7 dB in SNR for an input SNR of 0 dB. Varner et al. [1983] adopted a similar approach but replaced the pitch detector with an adaptive pitch tracker based on a 3-tap adaptive filter with adaptation in the z-plane. Kim and Un [1986] used a negative delay $D = -8$ with a filter length of 17 samples. They set α according to whether they wished to remove broadband ($\alpha = 0$) or periodic ($\alpha = 1$) noise and use a VAD to determine when to adapt the coefficients. Kawamura et al. [2005] uses $0 \leq \alpha \leq 0.5$ and describe a method for adaptively choosing the step size of the NLMS algorithm. They also suggest that improved enhancement can be obtained by cascading two adaptive filters.

Hoya et al. [1998] aim to remove narrow-band (periodic) noise and accordingly set $\alpha = 1$. They too use a VAD to detect the presence of speech, but in their case, they permit adaptation only when speech is absent. They observe that the adaptive filter can give a high gain at frequencies where $x(n)$ does not contain any correlated components. They therefore adopted an indirect approach in which a conventional two-channel adaptive filter was used with $x(n)$ as the interfering noise and a white noise signal as the desired response. The adapted filter coefficients were then transferred periodically to a second filter which was applied to the input signal $x(n)$. Other authors have suggested the use of "leaky" NLMS to overcome this problem.

Magotra et al. [1993] differ from other authors in using the filter coefficients $w(i) = g_i L^{-1} r_{0,0}^{-1}$ where L is the length of the filter. They claim that this results in much more rapid convergence than with NLMS albeit not to the optimum filter.

In order to suppress both periodic and broadband noise, Sasaoka et al. [2005] propose a cascade of three adaptive filters with different parameters (see Fig. 5). The first, with a long adaptation time-constant identifies stationary periodic noise such as arise from hum or machinery, the second identifies the periodic components of voiced speech while the final filter subtracts both periodic and broadband noise from the input signal while retaining the voiced periodic components. A slightly modified system is described in Sasaoka et al. [2007] which uses a variable step size to reduce the adaptation rate of the final filter when speech is present.

Ltd [2006] describes the adaptive filtering available in the Cedar Cambridge System (see Sec. **??**).

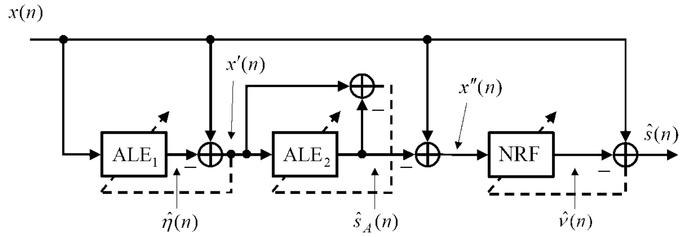

Figure 5: Three cascaded adaptive filters used to distinguish between periodic noise and voiced speech [Sasaoka et al., 2005]

They normally take the decorrelated output mixed with the original signal ($0.5 \leq \alpha \leq 1$) and allow the user to adjust the filter length, the "attack time" (adaptation rate) and "release time" (presumed to be the leaky parameter).

3.6 Model Based

Model-based speech enhancement uses prior knowledge in the form of an explicit stochastic model of speech and, in some cases, of the interferring noise. A number of different speech models are available including some combination of autoregressive (AR) models, cepstral coefficient models, hidden Markov models and pitch track models. A review of model-based techniques is given in Ephraim [1992].

Gibson et al. [1991] used a Kalman filter to enhance speech using an autoregressive model for both speech and noise. The noise model was assumed to be known in advance, while the coefficients of the speech model were estimated from the noisy speech during the enhancement procedure. The authors claimed that the use of an AR model for the noise resulted in significantly better performance than earlier approaches that assumed the noise to be white.

Yasmin et al. [1999] evaluate enhancement of voiced phonemes using an AR speech model with one of three different excitation signals: white noise, an impulse train and the LF glottal model [Fant et al., 1985]. They used a Kalman filter to enhance speech with added white noise at 5 dB SNR and found that the LF model gave the best performance with an SNR gain over the white noise model of about 1.3 dB. This differs from most model-based enhancement systems in the inclusion of an explicit voice source and therefore requires explicit pitch tracking which is performed using the LPC residual signal. As an alternative to Kalman Filtering, Shen and Deng [1999] use an H^∞ filter which does not require an explicit noise model but instead minimizes the worst possible effects of noise on the signal for any given noise energy. The authors found that this approach conssistently outperformed the use of a Kalman filter.

Attias and Deng [2001] model both speech and noise as a mixture of LPC spectra. They present an iterative EM procedure that estimates the LPC noise spectrum, the channel impulse response and the time-domain clean speech signal. The latter is effectively obtained by a weighted average of state-dependent Wiener filters where the weights correspond to mixture probabilities. Attias et al. [2001] presents a similar approach but uses a pretrained mixture model for the noise which simplifies the enhancement procedure considerably.

Enhancement methods based on an AR model of speech generally place no constraint other than stability on the estimated set of AR coefficients. In speech coding applications however, strong constraints are invariably placed on the permitted coefficient values by transforming them into the LSP domain before quantization [ITU-T, 1993a]. In an enhancement method described in Hansen and Clements [1991] and developed inPellom and Hansen [1998], the AR coefficients are transformed into the LSP domain and smoothed them across time using a triangular window. In addition, spectral constraints are introduced within each frame.

Ephraim et al. [1989a] use an HMM to model the speech but represents each state's output distribution using a mixture of LPC spectra rather than the conventional MFCC coefficients. The paper presents an approximate method of training the states which in Ephraim et al. [1989b] is supplemented by an exact method which gives a small improvement in performance.

Deng et al. [2004] perform enhancement in the log spectral domain using a Gaussian mixture speech model that incorporates both static and delta (time-derivative) parameters; these are assumed to be independent within each mixture and to have diagonal covariance matrices in the cepstral domain. The noise is estimated recursively using the methods described in Deng et al. [2001]. The authors claim that the short term temporal correlations captured by the use of delta coefficients are much more important than the longer term temporal correlations implied by the use of a Hiddden Markov Model.

Kristjansson and Hershey [2003] use Gaussian mixture models in the log spectral domain for both speech and noise using a large number (128) of frequency bins for an 8 kHz sample rate. The speech model comprised either 512 or 1024 mixtures while the noise model had only a single component. They reported that their technique gave enhanced speech of "exceptional quality" with an improvement in segmental SNR of 7 to 4 dB over the SNR range -5 to +15 dB SNR with no noticeable speech distortion at high SNR values.

3.7 Harmonic and Impulsive Noise

In this section we consider noise sources that are specifically bounded in the time-frequency domain. Interference that is temporally short can generally be considered impulsive, noises that are localized in frequency (whether individually or as a set of related components) can generally be considered harmonic. It is often possible to identify and clean these types of noise because they are well localized and the majority of the time-frequency representation of the signal is undamaged. Removing such interference can improve intelligibility because temporal and spectral masking in human hearing means that nearby regions of the time frequency plane will be inaudible despite being themselves uncorrupted.

Harmonic noise sources are usually heard as hums and buzzes which exhibit strong periodic components and corresponding harmonic structure. Suppression of harmonic noise that is temporally stationary usually begins with identifying the fundamental frequency. This can be achieved by manual spectral measurement or, in principle, by automatic methods. Notch filtering can then be applied both to the fundamental and to its harmonics to achieve suppression. When harmonic noise is time varying, it is necessary to track the fundamental and a more sophisticated approach, such as single channel adaptive filtering (Sec. 3.5), is required. Hum and buzz removal is a standard component of many audio tools including, for example, Ltd [2005] and M-Audio [2007].

Examples of impulsive noise are most commonly clicks, pops and crackles. Such noise may result from a variety of sources including impulsive radio frequency interference, faulty electrical connections and digital recording errors. Methods for identifying and cleaning impulsive noise are extensively discussed in Godsill and Rayner [1998] and the references therein. The common techniques usually require manual localisation of the impulsive noise. With knowledge of the time of occurrence of impulse noise, methods can then be used to interpolate the undamaged signal that exists before and after the noise in order to remove it. Pure interpolation methods assume that no useful information remains at the instant of the impulsive noise. Other methods assume that the original data may still be present during the impulsive noise and attempt to model it in order to achieve better accuracy. The statistical distribution of the samples, whether known, assumed or deduced, can be used to condition the interpolation processing. Autoregressive interpolation assumes the data can be well modelled by an autoregressive model and estimates the parameters of the autoregressive model using, for example, least squares error minimisation. Pitch information may also be employed to constrain the interpolated speech further.

3.8 Missing Feature Estimation

Harmonic and impulsive noise (see Sec. 3.7) can readily be identified and deleted in the time-frequency plane. Missing feature methods aim to identify arbitrarily shaped regions of the time-frequency plane that are dominated by noise and either remove or reconstruct them. The segmentation of the time-frequency plane into distinct sound sources is known as Computational Auditory Scene Analysis (CASA) and a good overview of the field is given in Wang and Brown [2006]. For CASA-based enhancement, the key step is the creation of a mask that identifies the noise-dominated regions of the time-frequency plane having an SNR below a suitable threshold, typically 0 to -5 dB. Some authors use a "soft" mask whose value denotes the probability (or sometimes the degree) of noise-domination. The effect of masking on human intelligibility was studied in Brungart et al. [2006]; they suggest an SNR threshold of 0 dB for a binary mask.

In Hu and Wang [2004] the input speech is transformed into the time-frequency domain by an auditory filterbank. Within each time-frequency "pixel" a dominant pitch is determined by searching for an autocorrelation peak. Compatible contiguous pixels are then grouped into segments and the segments are tentatively assigned to the foreground or background stream on the basis of their dominant pitch consistency. A pitch contour is now determined for the foreground stream and the foreground/background assignments are re-evaluated on the basis of compatibility with the pitch contour. Li et al. [2006] extends this approach by using an objective speech quality measure, P.563 [ITU, 2004b], to determine which pixels to include in the foreground stream.

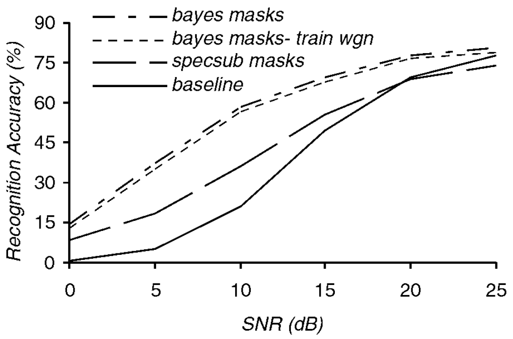

Figure 6: Recognition accuracy for speech corrupted with factory noise when noise corrupted portions of the time-frequency plane are reconstructed [Seltzer et al., 2004].

Seltzer et al. [2004] uses a Bayesian classifier to identify missing regions by means of a seven element feature vector which, for each of 20 Mel-frequency subbands, comprises:

1. the log ratio of the energy from a comb filter at the pitch period and that from the same filter but shifted by half the pitch frequency. The intention is to determine what fraction of the subband energy is contained in the periodic component. The pitch is determined using the RAPT pitch detector from Talkin [1995] which combines an autocorrelation method with dynamic programming.
2. the ratio of the highest to the second highest peak in the autocorrelation
3. the ratio of the subband energy to the total energy
4. the kurtosis of the subband signal. The assumption is that clean speech is super-Gaussian (i.e. has a higher kurtosis than a Gaussian) but that noisy speech will, because of the central limit theorem, have a lower kurtosis.
5. the ratio of the subband energy to the subband noise floor
6. the spectral flatness in neighbouring subbands.
7. the subband SNR estimated from spectral subtraction (Sec. 3.2).

The procedure requires a pitch detector which, since the first two features are used only for voiced frames, also serves as a voicing detector. Separate classifiers are trained for voiced and voiceless speech in each subband with a "noise pixel" defined as one with worse than -5 dB SNR. Clean speech spectra are reconstructed for the missing regions using cluster-based reconstruction [Raj et al., 2004] in which clean speech log spectra are modelled by a Gaussian mixture distribution. The parameters of each mixture are used to calculate a maximum likelihood estimate of the missing data conditioned on the observed non-masked bins and on it not exceeding the masked bins. The final estimate is then formed by weighting each of these individual estimates by the likelihood of the corresponding mixture. A second, correlation-based, reconstruction method was described in Raj et al. [2004] which determined correlations between pixels in the spectrogram and uses these correlations in time and frequency to estimate the missing regions. In their evaluation tests however, this second reconstruction method performed consistently worse than the cluster-based method.

The procedure was evaluated using an automatic speech recogniser and, as shown in Fig. 6, improved recognition performance at an SNR of 10 dB, from 15% to 60% for factory noise. The improvement was somewhat better for white noise and significantly worse for interference consisting of music. As can be seen from the figure, the performance is significantly better than using spectral subtraction to identify the noisy regions of the time-frequency plane and decreases only slightly when white noise, rather than matched noise, is used to train the classifier. It also decreased for both higher and lower SNR levels. Kim and Stern [2006] use a similar approach but claim that using a wider variety of noises in training improves performance.

3.9 Restoration of clipped signals

In the absence of other noise sources, clipped bandlimited signals are subject to strong constraints and their use in signal restoration is discussed in Abel and Smith [1991]. Phase distortion subsequent to clipping makes the problem more difficult because the clipped signal samples are harder to identify and the amplitude constraints no longer apply.

3.10 Bandwidth Expansion

In narrowband telephony, speech is limited to the frequency range of 0.3 to 3.4 kHz. The objective of artificial bandwidth expansion is to create a wideband version of the bandlimited speech signal by adding artificial low and high frequency components. It is believed that this can improve intelligibility to some extent but mostly that it will improve a listener's perceptual experience. A review of many existing methods is given in Iser and Schmidt [2005].

Several algorithms operate on the AR model of the speech signal in (7) where the problem is divided into extension of the spectral envelope, characterised by **a**, and extension of the excitation signal, $e_s(n)$ [Jax and Vary, 2003]. The extension is considered in two distinct regions a high band of 3.4 to 8 kHz and a low band in the region 50 to 300 Hz.

$$s(n) = -\mathbf{a}^T \mathbf{s}(n-1) + e_s(n), \tag{7}$$

In Cheng et al. [1994] an algorithm is developed to extend the high band of the speech signal. A recovery function based on the statistical dependence between the narrow band and the high band speech is developed. A vector quantisation approach was studied in Enbom and Kleijn [1999] where a code book is implemented with corresponding narrow-band and wide-band spectra using the MFCCs as features. The wide-band vector with the smallest error calculated with the narrow-band spectra is selected. This also aims to reconstruct only the high-band region.

3.11 Reverberation

Reverberation reduction methods generally fall into two major categories: speech model enhancement and blind system identification and equalisation where the former has had most success for single-microphone applications.

3.11.1 Speech model enhancement

An early technique in the class of speech enhancement dereverberation, that has not been of recent interest, was described in Oppenheim et al. [1968] and Oppenheim and Schafer [1975]. The authors observe that simple echoes are result in distinct peaks in the cepstrum of the speech signal. Consequently, they use a peak picking algorithm to identify such peaks and attenuate them. They also consider an alternative approach in which a low-pass weighting function is applied to the cepstrum on the assumption that most of the speech energy lies in the lower quefrencies. However, this approach was not found suitable for more complex reverberation models [Oppenheim et al., 1968].

A class of techniques emerged from the observation that the residual signal following linear prediction analysis contains peaks corresponding to the excitation events in voiced speech together with additional peaks due to the reverberant channel. Several methods for processing the LPC residual have subsequently been developed. These aim to suppress the additional peaks due to reverberation without degrading the original components of the residual in order that dereverberated speech can be synthesised using the processed residual and the all-pole filter resulting from prediction analysis on the reverberant speech. It is assumed that the effect of reverberation on the AR coefficients is negligible Gaubitch et al. [2006], Griebel and Brandstein [1999, 2001].

Yegnanarayana and Murthy [2000] provided a comprehensive study of the prediction residual of reverberant speech. They demonstrate that reverberation affects the prediction residual differently in different speech segments, depending on the energy in the signal and whether a segment is voiced or unvoiced. Motivated by these observations, the authors proposed using a regional weighting function based on the signal-to-reverberant ratio in each region together with a global weighting function derived from the short term signal energy. For the derivation of the SRR based weightings, the entropy function and the normalised error were used.

An adaptive algorithm was proposed in Gillespie et al. [2001] using a kurtosis maximising subband adaptive filter. The authors demonstrate that the kurtosis of the prediction residual decreases as a function of increased reverberation, which was also suggested in Yegnanarayana and Murthy [2000]. They use this observation to derive an adaptive filter that maximises the kurtosis of the prediction residual. The filter is applied directly to the observed signal rather than to the prediction residual and thus avoids the the need for an LPC synthesis stage. The adaptive filter is implemented in a multi-channel subband framework for increased efficiency. This approach is an example of a hybrid method between speech enhancement and blind system identification and inversion. An extension to this method was presented in Wu and Wang [2006], where it was combined with spectral subtraction to remove residual reverberation due to late reflections.

In Gaubitch and Naylor [2007] a spatio-temporal averaging method is described. The peaks in the LPC residual due to glottal closure instances are identified and enhanced glottal cycles are obtained by temporal averaging of neighbouring cycles. Each processed glottal cycle is used to update an equalisation filter that is applied during unvoiced speech regions. The LPC based methods described above all achieve moderate reverberation reduction when a single microphone is used.

A related method, proposed by Nakatani et al. [2005], assumes a sinusoidal speech model. First the fundamental frequency of the speech signal is identified from the reverberant observations, then follows the identification of the remaining sinusoidal components. Using the magnitudes and phases of these sinusoids, an enhanced speech signal is synthesised. Subsequently, the reverberant and the dereverberated speech signals are used to derive an equivalent equalisation filter. The processing is performed in short frames and the inverse filter is updated in each frame. It is shown that this inverse filter does tend to the RIR equalisation filter, however, it is very long and takes over an hour of training.

Spectral subtraction for dereverberation using a statistical model of the room impulse response was proposed in Lebart et al. [2001]. Further developments of this method, including extension to multiple microphones and generalisation of the room impulse response model, were done in Habets [2007].

3.11.2 Blind system identification and equalisation

In blind system identification, an estimate of the acoustic impulse response is first obtained from the speech signal affected by convolutive distortion which is then used to design an equalising filter. The blind identification process for a single microphone relies on the assumption that the channel varies much more slowly than the speech signal. In Hopgood and Rayner [2003] an all-pole model of the channel impulse response is used in contrast to the more commonly used FIR model. The channel impulse response is assumed stationary while the source signal is assumed to be a locally stationary, but globally nonstationary, AR process. In this way, the all-pole channel parameters can be identified by observing several frames of the input signal and collecting information about the poles either by using a histogram approach or a more robust Bayesian probabilistic framework. Over several frames, the poles due to the stationary channel become apparent and the channel can thus be identified. One major advantage of this method is that, by using an AR model of the channel, the order of the channel is reduced compared to the FIR channel models. A similar approach for FIR channels was proposed in Pacheco and Seara [2005]. A major problems with blind system identification methods remains their sensitivity to model order estimation errors and to additive noise.

If the impulse response, $h(n)$, is available, for example, from a blind system identification algorithm, equalisation can be achieved by an inverse system, $G(z)$, satisfying $G(z)H(z) = \kappa z^{-\tau}$, where κ and τ are arbitrary scale and delay factors. However, direct inversion of an acoustic channel is normally not feasible since: (i) it can be several thousand taps in length depending on the acoustic characteristics of the recording environment and the sampling frequency, (ii) it is generally non-minimum phase [Neely and Allen, 1979] and (iii) it may contain spectral nulls that after inversion give strong peaks in the spectrum resulting in narrow band noise amplification. Alternative approaches have been studied for single channel inversion. For example, single channel least squares inverse filters [Mourjopoulos et al., 1982, Mourjopoulos, 1994] can be designed by solving the optimisation problem

$$\hat{G}(z) = \arg\min_{G(z)} \left\| G(z)H(z) - \kappa z^{-\tau} \right\|^2 \tag{8}$$

This requires extremely long inverse filters and results in large processing delay. Homomorphic inverse filtering has been investigated [Neely and Allen, 1979], where the impulse response is decomposed into a minimum phase component, $H_{mp}(z)$ and an allpass component, $H_{ap}(z)$, such that $H(z) = H_{mp}(z)H_{ap}(z)$. Consequently, magnitude and phase are equalised separately, where an exact inverse can be found for the magnitude, while the phase can be equalised using, for example, matched filtering [**?**]. In a comparative study, Mourjopoulos concluded that the homomorphic approach can give more accurate results than the least squares method [Mourjopoulos et al., 1982], however, the decomposition of the impulse response into a minimum phase and allpass components is problematic and computationally expensive in practice and therefore the least squares approach is usually preferred. An important result is that audible distortions occur in the processed speech signal if equalisation is performed only on the magnitude response [**?**Neely and Allen, 1979]. The equalisation methods described above approximate an infinite support system with a finite length filter and consequently, only approximate equalisation can be achieved.

4 Noise Estimation

In many speech enhancement algorithms, the first step is to estimate the power spectrum of the noise. To do this, it is necessary to make use of prior knowledge about differences between the characteristics of noise and speech. Common assumptions are

1. that the short-time power spectrum of noise is more stationary than that of speech
2. that, within a narrow frequency band, the speech energy frequently falls to a low value
3. that the frequency of periodic noise sources changes very slowly with time; this is in contrast to voiced speech whose period changes more rapidly.

The estimation of the noise is almost always performed in a spectral or related domain for several reasons: speech and noise are partially separated in the spectral domain; spectral components of both speech and noise are somewhat decorrelated; psycho-acoustic models are conveniently applied in this domain. Thus the following domains, all possessing different advantages, are used: (i) complex spectral amplitudes, (ii) spectral magnitudes, (iii) spectral powers, (iv) log spectral powers, (v) Mel-spaced spectral amplitudes, (vi) Mel-spaced log spectral powers, (vii) Mel cepstral coefficients, (viii) AR coefficients. In each of these domains, the coefficients are most frequently taken to be Gaussian and uncorrelated; these assumptions are rarely well substantiated.

A well-presented evaluation of several noise spectrum estimation techniques is given in Ris and Dupont [2001] who found that the best performance of the tested algorithms was given by a combination of minimum statistics and harmonic filtering.

4.1 Noise Models

4.1.1 Stationary Spectral Model

The most common model of the noise is that it is a Gaussian process with a slowly changing power spectrum. The spectrum is normally represented by individual spectral, mel-spectral or cepstral coefficients. In Ephraim et al. [1989a] an all-pole spectrum model is used for the noise process and an approximation to the maximum likelihood estimate of the AR parameters is provided by conventional LPC analysis.

4.1.2 Two-component model

Rennie et al. [2006] differ from most authors in modelling the noise as the sum of a slowly evolving component and a random component. They claim that this model is both more realistic and allows better tracking of the evolving component. Both components are modelling in the Mel log power spectral domain. The power of the continuously evolving component, l_t, is modelled as a 1st order Gaussian AR process in each frequency bin, i.e. $(l_{t+1}|l_t) \sim N\left(l_{t+1}; l_t, \sigma_d^2\right)$ while the random component is zero-mean Gaussian with variance σ_n^2. To account for abrupt changes in noise level (e.g. opening a car window) there is a small but non-zero probability that the continuously evolving component reverts to its prior distribution which is a mixture of diagonal-covariance Gaussians; as an alternative to reverting to a fixed prior, they also suggest reverting to a minimum-statistics noise estimate. The paper gives update procedures for the mean and variance of the noise level components under the assumption of a fixed Gaussian mixture model for the speech. A similar noise model was also used implicitly by Manohar and Rao [2006] (see Sec. 3.3).

4.1.3 Hidden Markov Model

A number of authors present their noise model as a multi-state HMM, but in most cases they actually use only a single state (albeit with multiple mixture components). Graciarena and Franco [2003] review ways of estimating a noise HMM.

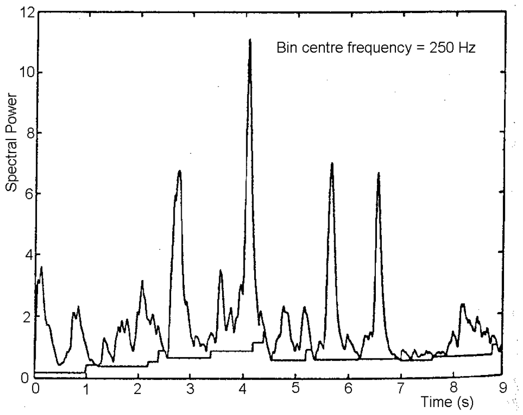

Figure 7: Power in the 250Hz subband of a noisy speech signal and the output of a minimum filter with $T = 0.8$ s [Martin, 1994]

4.2 Minimum Statistics

The use of minimum statistics for noise estimation was introduced in Martin [1994] and extended in Martin [2001]. The assumption is that in any frequency bin there will be times when there is little speech energy and that the energy will then be dominated by the noise. If we assume that these times occur at least once per time T, we can estimate the noise power as the minimum power that has arisen within the past T (typically 0.5 to 1.5 seconds). This is illustrated in Fig. 7 in which the upper trace shows the power in one subband (centred at 250 Hz) of a noisy speech signal. The lower trace in Fig. 7 shows the output of a minimum filter with $T = 0.8$ s. The output of the minimum filter will inevitably underestimate the true noise and it is necessary to compensate for this bias. In Martin [2001] the fixed compensation factor used in the original algorithm was replaced with factor that varied with time and frequency.

A similar approach was used in Doblinger [1995] but, instead of taking the minimum over T, the noise speech spectrum is smoothed using two different time constants; a short time constant is used when the energy in a frequency bin is decreasing to ensure rapid adaptation to a new minimum while a long time constant is used when the power increases to prevent adaptation to the speech energy. The approach is computationally efficient but is considered to perform less well than the minimum statistics approach because selecting the long time constant is a compromise between the response to sudden increases in noise and preventing the speech power from modulating the estimated noise power.

Stahl et al. [2000] noted that the use of the minimum makes the technique sensitive to outliers and investigated the use of other quantiles instead. They came to the conclusion that the median gave the best results when evaluated using a speech recogniser. Few people, however, appear to have followed up this work although Manohar and Rao [2006] demonstrates that it performs poorly on non-stationary noise.

4.3 Voice Activity Detectors (VADs)

A straightforward way to estimate the noise spectrum is to use a voice activity detector (VAD) to identify when speech is absent and to average the signal powers spectrum during these intervals. The appropriate averaging time-constant depends on the assumed stationarity of the noise. Many different VAD methods have been used.

4.3.1 Energy Histogram

McAulay and Malpass [1980] proposes a modification of an idea, attributed to Roberts [1978], which is based on the bimodality of the signal energy histogram taken over a 4 second window. Their algorithm determines an adaptive energy threshold to decide on the presence of speech but also includes fixed upper and lower thresholds which take priority. The adaptive threshold is chosen to lie at the 80th centile of the histogram of energies that are below the upper fixed threshold. This approach is modified in Compernolle [1989] which fits a 2-component Gaussian mixture model to the histogram of log energy and assumes that the lower component represents the noise. A similar approach is used by Hirsch and Ehrlicher [1995] in which an adaptive threshold is used in each frequency bin to eliminate speech frames and the peak of

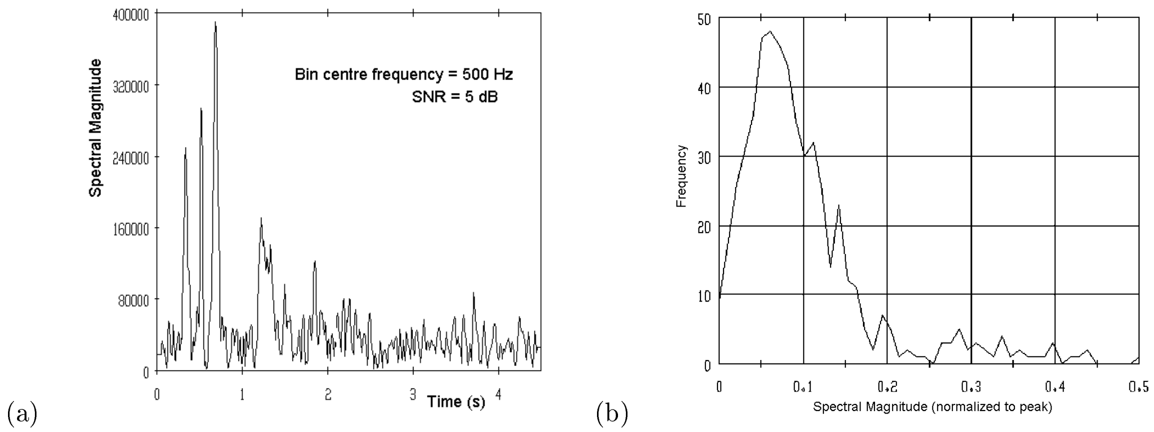

Figure 8: (a) Spectral magnitude envelope and (b) histogram of the 500 Hz subband of a noisy speech signal with SNR = 5 dB [Hirsch, 1993]

the histogram of recent noise frames is used as an estimate of the noise power in that bin. The reported accuracy of this approach, which is an extension of Hirsch [1993], is much greater than that of the VAD approach. The approach is illustrated in Fig. 8 which shows (a) the spectral magnitude envelope of the 500 Hz subband of a 4 second speech segment that has been corrupted with filtered Gaussian noise and (b) histogram of the values that it takes. The highest peak in the histogram arises from the noise whereas the much smaller peaks at higher magnitudes arise from the speech.

Freeman et al. [1989] describe the VAD used for GSM phones; it effectively compares the SNR averaged over all frequency bands to a threshold in order to perform the VAD decision. The same criterion was used in Yang [1993].

The NIST SPQA package [NIST, 1992] includes an algorithm to measure the noise level of the speech. It fits two raised cosine functions to the noisy speech log power histogram and assumes that the lower of the two corresponds to the noise. A similar approach is used in Dat et al. [2006] but instead a 2-component GMM is used. It is initialised using k-means clustering and then estimated using the EM algorithm.

Krubsack and Niederjohn [1994] observe that the energy in voiced speech is predominantly low frequency while that of unvoiced speech is often predominantly high frequency. They therefore use a voicing detector from Krubsack and Niederjohn [1991] to distinguish between the two and use different noise estimation algorithms for the two cases. For voiced speech, they use a comb filter at low frequencies to remove the speech and assume that speech energy is negligible at high frequencies. For unvoiced segments, they estimate only the low frequency noise spectrum.

4.3.2 Magnitude Histogram

Zhang and Gazor [2002] and Gazor and Zhang [2003b] demonstrate that the DFT or KLT coefficients of speech signals follow a Laplacian distribution rather than the more commonly assumed Gaussian or Gamma distributions. They therefore propose in Gazor and Zhang [2003a] a voice activity detector that models the noisy signal subband coefficients as the sum of zero-mean Laplacian and Gaussian random variables respectively. They find that using the KLT rather than the DFT gives marginally better performance.

4.3.3 Energy Waveform

Marzinzik and Kollmeier [2002] presents an algorithm for speech pause detection that uses the energy waveform in low, high and full frequency bands. The maximum and minimum values of these are tracked and used to decide when speech pauses are present. The performance of this VAD is compared to that used in G.729 [ITU-T, 1993a] and shown to perform better.

In Sugiyama et al. [2002], Kato et al. [2003] and Kato et al. [2006] a weighted noise estimation procedure used in 3G handsets is described. For each frame, a raw noise estimate is generated that is approximately equal to the the noisy speech divided by the estimated SNR. The smoothed noise estimate is then formed by averaging, in each frequency bin, the N most recent raw estimates for which the SNR was below a fixed threshold.

A VAD described in Annex A of ETSI Standard ES202050 [ETSI, 2007] identifies the increase in energy associated with the onset of voiced speech. It makes a VAD decision based on the values of three indicators: (a) a sudden increase in overall energy (b) a sudden increase in low frequency energy and (c) a sudden increase in the variance of energy across the spectrum.

4.3.4 Periodicity

Lin et al. [2007] detect voiced speech by looking for strong periodicity. They do this by assuming that the fundamental is a subharmonic of the largest spectral peak, summing all its aligned harmonics below 1.25 kHz and measuring how sharp the resultant peak is. Unvoiced speech is assumed to be possible for a 0.2 s hangover interval following a burst of voiced speech. This is preceded by the speech probability detector of Cohen and Berdugo [2002] to give a composite VAD which controls noise adaptation.

Lin and Goubran [2005] combine a periodicity detector based on the Fourier transform of the AMDF function with a modified minimum statistic approach to control noise spectrum averaging.

4.4 Soft Decision VAD

Rather than using a hard VAD decision, it is possible instead to use a soft decision so that the estimated noise spectrum is updated all the time but with a time-constant that varies according to the probability of speech presence, i.e. a long time constant is used when speech is absent. The soft-decision VAD that is used by Sohn and Sung [1998] is based on a likelihood ratio that is equivalent to the Itakura-Saito distortion measure or cross entropy between background noise and observed signal Shore [1981], Gray et al. [1981]. A similar approach is used in Malah et al. [1999] where the estimated SNR averaged across all frequencies is used to control adaptation together with an additional frequency-dependent factor that depends on the estimated speech presence in each frequency bin. A simpler version of this approach is in Lin et al. [2003a,b] where the noise adaptation time constant depends on a sigmoid function of the noisy speech to noise ratio.

In Cohen and Berdugo [2001, 2002], a speech probability estimate is determined by taking the ratio between the power in the current frame and its minimum within a specified time frame. This is then used to control which sections of the noisy speech are averaged to estimate the noise power. The authors claim that this gives an estimate with less bias and reduced variance than the original approach from Martin [1994]. An improved version of the noise estimator was given in Cohen [2003] which uses a two-iteration procedure that refines an initial speech presence detector. Rangachari et al. [2004] extends this approach in two ways: it uses a different way of calculating the minimum spectrum that has a lower latency (0.5 instead of 1.5 s) and a frequency-dependent threshold on the ratio of noisy speech spectrum to minimum spectrum which is used to estimate the “speech presence probability” and thence to control the adaptation rate. The algorithm is improved slightly in Rangachari and Loizou [2006] by smoothing the speech presence probability over time which implicitly accounts for its correlation between successive frames. The paper includes comparisons with Martin [2001], Hirsch and Ehrlicher [1995], Cohen and Berdugo [2002] and Cohen [2003].

4.5 Noise Estimation using a Speech Model

Joint estimation of speech and noise from a combined speech and noise model has been widely used in speech recognition in which the probability of a speech state is determined by marginalising over all possible noise states. The technique was introduced in Varga and Moore [1990] and extended by Gales [1995] and in subsequent papers [Gales and Young, 1993, 1995, 1996]. These authors used HMMs to model both speech and noise in the mel-cepstral domain giving a combined model whose state count was the product of the speech and noise model counts. In practice, the noise model normally had very few states and often only one. Modifications have subsequently been been made in [Chang et al., 1998, Chang and Chung, 1998]. Kristjansson et al. [2001] use a noise GMM; they found that selecting the max likelihood noise state performed similarly to marginalising over all noise states. A good introduction is given in Raju [2003].

When using this approach for noise estimation, the noise state may be estimated by marginalising over the speech states. This approach was adopted in Yao and Nakamura [2002], Yao and Lee [2003] where a particle filter is used to represent the possible sequences of speech states. In this application the speech model can be quite simple and the authors used only 18 states with 8 Gaussian mixtures per state in the log spectral domain. In a development of this work, Lee and Yao [2004] estimates the noise characteristics in the log spectral domain using EM but without a particle filter. A difficulty with the joint estimation approach when used for enhancement is that it is necessary to estimate the absolute speech energy; speech models developed for recognition generally ignore the overall speech level since it does not affect the speech state sequence. Subramanya et al. [2005] model speech using a 4 component GMM in the magnitude-normalised spectral domain rather than the more usual cepstral domain; this is the correct domain for adding noise and speech and its use avoids the difficulties that arise from the non-linear logarithmic transformation into the cepstral domain. They claim that applying magnitude normalisation significantly reduces the complexity required in the model although it entails modelling the overall speech energy separately.

Using a sinusoidal model of speech from McAulay and Quatieri [1986] and an assumption that the noise signal has zero kurtosis, Nemer et al. [1999] estimates the speech energy in each of fifty subbands from the kurtosis of the bandpass filtered signal.

In T. Kristjansson and Deng [2001] and Frey et al. [2001a] an EM approach is used to estimate the speech, noise and channel adaptively in the log spectrum domain. Each of these three components is represented with a Gaussian mixture model. In most of the examples they give, the noise model comprised only a single mixture but, for the case of aircraft noise at an airport, they investigated the use of up to 16 mixtures (the speech model, in contrast, used 256 mixtures). A 1st-order Taylor-series approximation is used to linearise the mapping between the log power domain and the linear power domain. In Frey et al. [2001b], the authors found that their adaptive noise modelling reduced speech recognition word errors by about 15% compared to a non-adaptive model estimated from the beginning of the recording and that increasing the noise model from 1 to 4 mixtures gave a further improvement of up to 0.3%. A similar model (in the Mel log spectral domain) is used in Deng et al. [2003b] who develop a recursive estimate of the parameters of the single-mixture noise model which was extended to a Bayesian formulation in Deng et al. [2003a]. A very similar, albeit non-iterative, approach is used in Afify and Siohan [2001, 2004].

5 Databases

5.1 Speech Enhancement Databases

There are a number of publicly or commercially available databases of noisy speech that may be of use for the evaluation of speech cleaning methods. The precise definition of SNR is not wholly uniform between databases, but most use the method described in ITU-T P.56 [ITU-T, 1993c] to define a measure of the active speech level that is unaffected by pauses in the speech activity. All the databases use 16 bit linear quantisation.

5.1.1 NOISEX

The NOISEX database [Varga et al., 1992, Varga and Steeneken, 1993] was created to provide standardised test material for speech recognition in noise and has been widely used. It contains speech recordings from two speakers (one male and one female) reading tables of isolated digits and digit triples of total duration 16 min per speaker. The database also includes eight noise recordings taken from a NATO database [Steeneken and Geurtsen, 1988] each of about 3.8 min duration. All data is recorded at 16 kHz sample rate. The noise signals are:

Speech Noise Broadband noise whose long-term spectrum matches that of speech

Machine Gun Repeated firing bursts from a 0.5 calibre machine gun

STI Test Signal A test signal for the measurement of the Speech Transmission Index as described in [Steeneken and Houtgast, 1980].

Lynx recording from the platform of a Lynx helicopter

F16 Cockpit noise recorded from the copilot seat of an F16 fighter under various flight conditions.

Car A Volvo 340 travelling on an asphalt road at 120 km/h

Factory Noise from a car factory including electrical welding

Operations Room recording from the operations room of a destroyer

5.1.2 TIMIT, NTIMIT and CTIMIT

The TIMIT database contains recordings from 630 speakers, each of whom speaks 10 sentences lasting a few seconds each. Two of the sentences are intended to distinguish between American dialects and are read by all speakers. About 70% of the speakers are male.

The NTIMIT database was created by playing the TIMIT recordings over telephone landlines to give speech that is realistically degraded [Jankowski et al., 1990]. In a similar way, the CTIMIT database was created by recording the TIMIT sentences over cellphone lines [Brown and George, 1995].

5.1.3 Aurora

The Aurora database [Hirsch and Pearce, 2000] contains extracts from the TIDIGITS database [Leonard, 1984] which have been filtered to telephone bandwidth and corrupted with added noise. The spoken sentences consist of digit strings containing between one and seven digits. Eight noises are used: subway, babble, car noise, exhibition hall, restaurant, street, airport and train station. The recordings are at 8 kHz and have all been filtered with a G.712 characteristic.

5.1.4 ITU-T Coded Speech Database

This database [ITU, 1998b], originally developed for evaluating the G.729 codec [ITU-T, 1993a], contains speech (short sentences) that is normalized for level and then degraded by passing it through the addition of acoustic noise, passage through one or more codecs and the introduction of channel errors. The database includes MOS ratings of the resultant degraded speech signals. The acoustic noise sources are white noise, office babble, two examples of car noise, background music and street noise.

5.1.5 Noizeus

The Noizeus database was created to evaluate speech enhancement algorithms and consists of spoken sentences to which has been added the eight noise recordings from the Aurora database (Sec. 5.1.3) at 0, 5, 10 and 15 dB SNR. There are thirty different sentences taken from Rothauser et al. [1969] with five spoken by each of six speakers. The recordings are at 8 kHz and have been filtered by a simulated telephone handset.

5.1.6 SpEAR Database

The Speech Enhancement and Assessment Resource [Wan et al., 1998] is under development at the Centre for Spoken Language Understanding (CSLU) at the Oregon Graduate Institute. It aims to create a speech database and toolkit for assessing speech enhancement algorithms. The database currently includes acoustically combined speech and noise and also “Lombard speech” recorded in a noisy environment.

5.1.7 RSPL Noises

The Robust Speech Processing Lab at Univ Colorado has made available 10 noise source recordings each lasting only about 4 seconds at 8 kHz sampling rate. The noises are: aircraft cockpit, babble, city rain, communications channel, helicopter fly-by, automobile highway, large city, large crowd, PC computer cooling fan, SUN computer cooling fan, white Gaussian noise [Hansen and Arslan, 1995].

5.2 Acoustic Impulse Responses

Acoustic impulse responses can be measured with a number of different acoustic sources: a true impulse (e.g. a gun shot), white noise, pseudo-random noise (MLS or IRS sequences), a linear or logarithmically swept sine wave. A discussion of the relative methods is given by Stan et al. [2002] who conclude that the MLS or IRS techniques perform best in the presence of non-white noise but that the logarithmic swept sine wave approach [Farina, 2007] is better in quiet environments. Ajdler et al. [2007] present a technique for measuring all impulse responses along a line by means of a moving microphone. A number of measured impulse responses are available and are described below; some use a single microphone while others use a 4-microphone recorder to retain directional information.

- Audioease [2007] manufacture the Altiverb plugin for convolutional reverberation and which includes a large number of recorded impulse resonses including cars, planes and rooms.

- Waves [2007] manufacture the IR Parametric Convolution plugin which includes a number of recorded impulse responses from concert halls, clubs, cars and other rooms.
- Noisevault [2007] provides measured impulse responses from a number of rooms and concert halls as well as from microphones and other recording equipment.

References

Subjective assessment of sound quality, 1990.

Subjective performance evaluation of telephone band and wideband codecs, 1998a.

ITU-T coded-speech database, February 1998b.

Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, February 2001.

Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms, November 2003.

Analysis and interpretation of INMD voice-service measurements, May 2004a.

Single-ended method for objective speech quality assessment in narrow-band telphony applications, 2004b.

J. S. Abel and J. O. Smith, III. Restoring a clipped signal. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, pages 1745–1748, 1991. doi: 10.1109/ICASSP.1991.150655.

M. Afify and O. Siohan. Sequential noise estimation with optimal forgetting for robust speech recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 229–232, 2001. doi: 10.1109/ICASSP.2001.940809.

M. Afify and O. Siohan. Sequential estimation with optimal forgetting for robust speech recognition. *IEEE Trans. Speech Audio Process.*, 12(1):19–26, 2004. doi: 10.1109/TSA.2003.819954.

Thibaut Ajdler, Luciano Sbaiz, and Martin Vetterli. Dynamic measurement of room impulse responses using a moving microphone. *J. Acoust. Soc. Am.*, 122:1636–1645, 2007. doi: 10.1121/1.2766776.

J. Allen and L. Radiner. A unified approach to short-time Fourier analysis and synthesis. *Proc. IEEE*, 65(11):1558–1564, 1977.

J. B. Allen. Short term spectral analysis, synthesis, and modification by discrete Fourier transform. *IEEE Trans. Acoust., Speech, Signal Process.*, 25(3):235–238, June 1977.

ANSI. Methods for the calculation of the articulation index. ANSI Standard ANSI S3.5–1969, American National Standards Institute, New York, 1969.

ANSI. Methods for the calculation of the speech intelligibility index. ANSI Standard S3.5–1997 (R2007), American National Standards Institute, 1997.

H. Attias and L. Deng. Speech denoising and dereverberation using probabilistic models. *Advances in Neural Information Processing Systems (NIPS)*, 13:758–764, 2001.

H. Attias, L. Deng, A. Acero, and J. Platt. A new method for speech denoising and robust speech recognition using probabilistic models for clean speech and for noise. In *Proc. European Conf. on Speech Communication and Technology*, pages 1903–1906, 2001.

Audioease. Altiverb plugin. Technical report, Audioease, 2007. URL `http://www.audioease.com/Pages/Altiverb/AltiverbMain.html`.

J. G. Beerends, P. Gray, A. P. Hekstra, and M. P. Hollier. Call for proposals for a single-ended speech quality assessment method for non-intrusive measurements on live voice traffic. Technical Report COM12-C11, ITU-T, 2000.

J. Benesty, S. Makino, and J. Chen, editors. *Speech Enhancement*. Springer, 2005.

M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 208–211, 1979.

R. Bilger, J. Nuetzel, W. Rabinowitz, and C. Rzeczkowski. Standardization of a test of speech perception in noise. *J. Speech. Hear. Res.*, 27:32–48, 1984.

S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-27(2):113–120, April 1979.

K. Brandenburg. Evaluation of quality for audio encoding at low bit rates. In *Proc. Audio Eng. Soc. Conventions*, number 2433, February 1987.

K. L. Brown and E. B. George. CTIMIT: a speech corpus for the cellular environment with applications to automatic speech recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 105–108, 1995. doi: 10.1109/ICASSP.1995.479284.

Douglas S. Brungart, Peter S. Chang, Brian D. Simpson, and DeLiang Wang. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. *J. Acoust. Soc. Am.*, 120:4007–4018, 2006.

O. Cappe. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Trans. Speech Audio Process.*, 2(2):345–349, April 1994. doi: 10.1109/89.279283.

Y. H. Chang and Y. J. Chung. Improved HMM parameter compensation method for noise-robust speech recognition using state-dependent association factor. *IEE Electronics Lett.*, 34(8):724–725, April 1998.

Y. H. Chang, Y. J. Chung, and S. U. Park. Improved model parameter compensation methods for noise-robust speech recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 561–564, May 1998. doi: 10.1109/ICASSP.1998.674492.

Bin Chen and P. C. Loizou. Speech enhancement using a MMSE short time spectral amplitude estimator with Laplacian speech modeling. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 1097–1100, March 2005. doi: 10.1109/ICASSP.2005.1415309.

Y. M. Cheng, D. O'Shaughnessy, and P. Mermelstein. Statistical recovery of wideband speech from narrowband speech. *IEEE Trans. Speech Audio Process.*, 2(4):544–548, October 1994. doi: 10.1109/89.326637.

C. Chernick, S. Leigh, K. Mills, and R. Toense. Testing the ability of speech recognizers to measure the effectiveness of encoding algorithms for digital speech transmission. In *Proc. Military Communications Conf*, volume 2, pages 1468–1472, October 1999. doi: 10.1109/MILCOM.1999.821447.

F. R. Clarke. Technique for evaluation of speech systems. Final Report AD 473–995, Stanford Research Institute, August 1965.

I. Cohen. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Trans. Speech Audio Process.*, 11(5):466–475, September 2003. doi: 10.1109/TSA.2003.811544.

I. Cohen. On the decision-directed estimation approach of ephraim and malah. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, May 2004. doi: 10.1109/ICASSP.2004.1325980.

I. Cohen and B. Berdugo. Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Signal Process. Lett.*, 9(1):12–15, January 2002. doi: 10.1109/97.988717.

Israel Cohen and Baruch Berdugo. Speech enhancement for non-stationary noise environments. *Signal Processing*, 81(11):2403–2418, November 2001. doi: 10.1016/S0165--1684(01)00128--1.

P. Combescure, A. Le Guyader, and A. Gilloire. Quality evaluation of 32 kbit/s coded speech by means of degradation category ratings. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 988–991, 1982.

Dirk Van Compernolle. Noise adaptation in a hidden Markov model speech recognition system. *Computer Speech and Language*, 3:151–167, 1989.

Tran Huy Dat, Kazuya Takeda, and Fumitada Itakura. On-line Gaussian mixture modeling in the log-power domain for signal-to-noise ratio estimation and speech enhancement. *Speech Communication*, 48(11):1515–1527, November 2006.

G. M. Davis, editor. *Noise Reduction in Speech Applications.* CRC Press, May 2002.

M. Dendrinos, S. Bakamidis, and G. Carayannis. Speech enhancement from noise: a regenerative approach. *Speech Communication*, 10(1):45–67, February 1991. doi: 10.1016/0167--6393(91)90027-Q.

Li Deng, J. Droppo, and A. Acero. Recursive noise estimation using iterative stochastic approximation for stereo-based robust speech recognition. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 81–84, December 2001.

Li Deng, J. Droppo, and A. Acero. Incremental Bayes learning with prior evolution for tracking nonstationary noise statistics from noisy speech data. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, April 2003a.

Li Deng, J. Droppo, and A. Acero. Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition. *IEEE Trans. Speech Audio Process.*, 11(6):568–580, November 2003b. doi: 10.1109/TSA.2003.818076.

Li Deng, J. Droppo, and A. Acero. Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features. *IEEE Trans. Speech Audio Process.*, 12(3):218–233, May 2004. doi: 10.1109/TSA.2003.822627.

E. D. Dickson and D. L. Chadwick. Speech audiometry in assessment of deafness. *J Laryngol Otol*, 64 (8):464–481, 1950.

D. Dirks, D. Morgan, and J. Dubno. A procedure for quantifying the effects of noise on speech recognition. *J. Speech. Hear. Disord.*, 47:114–123, 1982.

G. Doblinger. Computationally efficient speech enhancement by spectral minima tracking in subbands. In *Proc. European Conf. on Speech Communication and Technology*, pages 1513–1516, Madrid, September 1995.

J. Egan. Articulation testing methods. *Laryngoscope*, 58(9):955–991, 1948.

M. Elhilali, T. Chi, and S. A. Shamma. A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. *Speech Communication*, 41:331–348, 2003.

N. Enbom and W. B. Kleijn. Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficents. In *IEEE Workshop on Speech Coding*, pages 171–173, Porvoo, Finland, June 1999.

Y. Ephraim. Statistical-model-based speech enhancement systems. *Proc. IEEE*, 80(10):1526–1555, October 1992.

Y. Ephraim and I. Cohen. Recent advancements in speech enhancement. In R. C. Dorf, editor, *The Electrical Engineering Handbook, Circuits, Signals, and Speech and Image Processing*. CRC Press, third edition, 2006.

Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Process.*, 32(6):1109–1121, December 1984.

Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Process.*, 33(2):443–445, 1985.

Y. Ephraim and H. L. Van Trees. A signal subspace approach for speech enhancement. *IEEE Trans. Speech Audio Process.*, 3(4):251–266, July 1995. doi: 10.1109/89.397090.

Y. Ephraim, D. Malah, and B.-H. Juang. On the application of hidden Markov models for enhancing noisy speech. *IEEE Trans. Acoust., Speech, Signal Process.*, 37(12):1846–1856, December 1989a. doi: 10.1109/29.45532.

Y. Ephraim, D. Malah, and B.-H. Juang. Speech enhancement based upon hidden Markov modeling. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 353–356, May 1989b. doi: 10.1109/ICASSP.1989.266438.

Y. Ephraim, H. Lev-Ari, and W. J. J. Roberts. A brief survey of speech enhancement. In *The Electronic Handbook*. CRC Press, second edition, February 2005.

ETSI. Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms. ETSI Standard ES 202 050, ETSI, January 2007.

G. Fairbanks. Test of phonemic differentiation: the rhyme test. *J. Acoust. Soc. Am.*, 30(7):596–600, 1958.

G. Fant, J. Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 26(4):1–13, 1985.

Angelo Farina. Advancements in impulse response measurements by sine sweeps. In *Proc. AES Convention*, Vienna, May 2007.

H. Fletcher and R. H. Galt. The perception of speech and its relation to telephony. *J. Acoust. Soc. Am.*, 22(2):89–151, 1950.

H. Fletcher and J. Steinberg. Articulation testing methods. *Bell Syst. Tech. J.*, 8:806–854, 1929.

D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd. The voice activity detector for the pan-european digital cellular mobile telephone service. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 369–372, May 1989. doi: 10.1109/ICASSP.1989.266442.

N. R. French and J. C. Steinberg. Factors governing the intelligibility of speech sounds. *J. Acoust. Soc. Am.*, 19(1):90–119, 1947.

Brendan J. Frey, Li Deng, Alex Acero, and Trausti Kristjansson. ALGONQUIN: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition. In *Proc. European Conf. on Speech Communication and Technology*, Aalborg, September 2001a.

Brendan J. Frey, Trausti T. Kristjansson, Li Deng, and Alex Acero. Learning dynamic noise models from noisy speech for robust speech recognition. In *Proc. Neural Information Processing Conf*, 2001b.

Benjamin Friedlander. Lattice filters for adaptive processing. *Proc. IEEE*, 70(8):829–867, 1982. ISSN 0018–9219.

J. Friedman. Multivariate adaptive regression splines. *Ann. Stat.*, 19(1):1–67, 1991.

D. B. Fry. Word and sentence tests for use in speech audiometry. *Lancet*, 278(7195):197–199, July 1961. doi: 10.1016/S0140--6736(61)90480--9.

D. B. Fry and P. M. T. Kerridge. Tests for the hearing of speech by deaf people. *Lancet*, 233(6020): 106–109, January 1939. doi: 10.1016/S0140--6736(00)60050--8.

M. J. F. Gales. *Model-based Techniques for Noise Robust Speech Recognition.* PhD thesis, Cambridge University, 1995.

M. J. F. Gales and S. J. Young. Cepstral parameter compensation for HMM recognition in noise. *Speech Communication*, 12:231–239, July 1993.

M. J. F. Gales and S. J. Young. Robust speech recognition in additive and convolutional noise using parallel model combination. *Computer Speech and Language*, 9(4):289–307, October 1995.

M. J. F. Gales and S. J. Young. Robust continuous speech recognition using parallel model combination. *IEEE Trans. Speech Audio Process.*, 4:352–359, September 1996. doi: 10.1109/89.536929.

N. D. Gaubitch and P. A. Naylor. Spatiotemporal averaging method for enhancement of reverberant speech. In *Proc. IEEE Intl. Conf. Digital Signal Processing (DSP)*, Cardiff, UK, July 2007. doi: 10.1109/ICDSP.2007.4288655.

N. D. Gaubitch, D. B. Ward, and P. A. Naylor. Statistical analysis of the autoregressive modeling of reverberant speech. *J. Acoust. Soc. Am.*, 120(6):4031–4039, December 2006.

S. Gazor and Wei Zhang. A soft voice activity detector based on a Laplacian–Gaussian model. *IEEE Trans. Speech Audio Process.*, 11(5):498–505, September 2003a. doi: 10.1109/TSA.2003.815518.

S. Gazor and Wei Zhang. Speech probability distribution. *IEEE Signal Process. Lett.*, 10(7):204–207, July 2003b. doi: 10.1109/LSP.2003.813679.

J. D. Gibson, B. Koo, and S. D. Gray. Filtering of colored noise for speech enhancement and coding. *IEEE Trans. Signal Process.*, 39(8):1732–1742, 1991. doi: 10.1109/78.91144.

B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio. Speech dereverberation via maximum-kurtosis subband adaptive filtering. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 6, pages 3701–3704, 2001.

S. J. Godsill and P. J. W. Rayner. *Digital Audio Restoration: A Statistical Model Based Approach.* Springer, 1998.

M. Graciarena and H. Franco. Unsupervised noise model estimation for model-based robust speech recognition. In *Proc. ASRU IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 351–356, December 2003. doi: 10.1109/ASRU.2003.1318466.

V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn. Low-complexity, nonintrusive speech quality assessment. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(6):1948–1956, November 2006. doi: 10.1109/TASL.2006.883250.

P. Gray, M. P. Hollier, and R. Massara. Non-intrusive speech quality assessment using vocal tract models. *IEE Proc. Vision Image Signal Processing*, 147(6):493–501, 2000. doi: 10.1049/ip-vis:20000539.

R. Gray, A. Gray, G. Rebolledo, and J. Shore. Rate-distortion speech coding with a minimum discrimination information distortion measure. *IEEE Trans. Inf. Theory*, 27(6):708–721, November 1981.

S. M. Griebel and M. Brandstein. Microphone array speech dereverberation using coarse channel modeling. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 201–204, November 2001.

S. M. Griebel and M. S. Brandstein. Wavelet transform extrema clustering for multi-channel speech dereverberation. In *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, pages 52–55, Pocono Manor, Pennsylvania, September 1999.

E. A. P. Habets. *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement.* PhD thesis, Technische Universiteit Eindhoven, 2007. URL `http://alexandria.tue.nl/extra2/200710970.pdf`.

M. Haggard and I. Mattingly. A simple program for synthesizing British English. *IEEE Trans. Audio Electroacoust.*, 16:95–99, 1968.

J. Hansen and B. Pellom. An effective quality evaluation protocol for speech enhancement algorithms. In *Proc. Intl. Conf. on Spoken Lang. Processing (ICSLP)*, volume 7, pages 2819–2822, 1998.

J. H. L. Hansen. Speech enhancement employing adaptive boundary detection and morphological based spectral constraints. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 901–904, Toronto, April 1991. doi: 10.1109/ICASSP.1991.150485.

J. H. L. Hansen and L. M. Arslan. Robust feature-estimation and objective quality assessment for noisy speech recognition using the credit card corpus. *IEEE Trans. Speech Audio Process.*, 3(3):169–184, 1995. doi: 10.1109/89.388143.

J. H. L. Hansen and M. A. Clements. Constrained iterative speech enhancement with application to speech recognition. *IEEE Trans. Signal Process.*, 39(4):795–805, April 1991. doi: 10.1109/78.80901.

J. H. L. Hansen, V. Radhakrishnan, and K. H. Arehart. Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(6):2049–2063, November 2006. doi: 10.1109/TASL.2006.876883.

Peter S. K. Hansen. *Signal Subspace Methods for Speech Enhancement.* PhD thesis, Lyngby, September 1997.

S. Haykin. *Adaptive Filter Theory.* Prentice-Hall, fourth edition, 2002.

M. Hecker and C. Williams. Choice of reference conditions for speech peference tests. *J. Acoust. Soc. Am.*, 39(5):946–952, 1966.

R. Hegger, H. Kantz, and L. Matassini. Noise reduction for human speech signals by local projections in embedding spaces. *IEEE Trans. Circuits Syst. I*, 48(12):1454–1461, December 2001. doi: 10.1109/TCSI.2001.972852.

H. G. Hirsch and C. Ehrlicher. Noise estimation techniques for robust speech recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 153–156, 1995.

Hans-Gunter Hirsch. Estimation of noise spectrum and its application to SNR-estimation and speech enhancement. Technical Report TR-93–012, ICSI Berkeley, Berkeley, 1993. URL `citeseer.ist.psu.edu/hirsch93estimation.html`.

Hans-Gunter Hirsch and David Pearce. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. ISCA Workshop on Automatic Speech Recognition*, pages 181–188, Paris, September 2000.

Inga Holube and Birger Kollmeier. Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. *J. Acoust. Soc. Am.*, 100(3):1703–1716, September 1996.

J. R. Hopgood and P. J. W. Rayner. Blind single channel deconvolution using nonstationary signal processing. *IEEE Trans. Speech Audio Process.*, 11(5):476 – 488, September 2003.

A. House, C. Williams, M. Hecker, and K. Kryter. Articulation testing methods: consonant differentiation with a closed response set. *J. Acoust. Soc. Am.*, 37(1):158–166, 1965.

T. Houtgast and H. J. M. Steeneken. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *J. Acoust. Soc. Am.*, 77(3):1069–1077, 1985. doi: 10.1121/1.392224.

T. Hoya, J. A. Chambers, N. Forsyth., and P. A. Naylor. Steady-state solutions of the extended LMS algorithm for stereophonic acoustic echo cancellation. In *Proc. European Signal Processing Conf. (EUSIPCO)*, pages 977–980, 1998.

G. Hu and D. L. Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Trans. Neural Netw.*, 15(5):1135–1150, 2004. doi: 10.1109/TNN.2004.832812.

Y. Hu and P. C. Loizou. Evaluation of objective measures for speech enhancement. In *Proc. Interspeech Conf.*, pages 1447–1450, 2006.

Yi Hu and P. C. Loizou. A subspace approach for enhancing speech corrupted by colored noise. *IEEE Signal Process. Lett.*, 9(7):204–206, July 2002a. doi: 10.1109/LSP.2002.801721.

Yi Hu and P. C. Loizou. A subspace approach for enhancing speech corrupted by colored noise. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 573–576, May 2002b.

Yi Hu and P. C. Loizou. A perceptually motivated approach for speech enhancement. *IEEE Trans. Speech Audio Process.*, 11(5):457–465, 2003. ISSN 1063–6676. doi: 10.1109/TSA.2003.815936.

Yi Hu and P. C. Loizou. Subjective comparison and evaluation of speech enhancement algorithms. *Speech Communication*, 49(7–8):588–601, July 2007. doi: 10.1016/j.specom.2006.12.006.

B. Iser and G. Schmidt. Bandwidth extension of telephony speech. *EURASIP Newsletter*, 16(2):2–24, June 2005.

ITU-T. Coding of speech at 8 kbit/s using conjugate-structure alebraic-code-excited line-prediction (CS-ACELP), March 1993a.

ITU-T. Effect of transmission impairments, March 1993b.

ITU-T. Objective measurement of active speech level, March 1993c.

C. R. Jankowski, Jr., , A. Kalyanswamy, S. Basson, and J. Spitz. NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 109–112, 1990. doi: 10.1109/ICASSP.1990.115550.

P. Jax and P. Vary. On artificial bandwidth extension of telephone speech. *Signal Processing*, 83:1707–1719, 2003.

S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen. Reduction of broad-band noise in speech by truncated QSVD. *IEEE Trans. Speech Audio Process.*, 3(6):439–448, November 1995. doi: 10.1109/89.482211.

M. T. Johnson, A. C. Lindgren, R. J. Povinelli, and Xiaolong Yuan. Performance of nonlinear speech enhancement using phase space reconstruction. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, April 2003.

D. Kalikow, K. Stevens, and L. Elliott. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *J. Acoust. Soc. Am.*, 61(5):1337–1351, 1977.

M. Kato, A. Sugiyama, and M. Serizawa. A family of 3GPP-standard noise suppressors for the AMR codec and the evaluation results. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, April 2003.

Masanori Kato, Akihiko Sugiyama, and Masahiro Serizawa. Noise suppression with high speech quality based on weighted noise estimation and MMSE STSA. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, 89(2):43–53, 2006.

A. Kawamura, Y. Iiguni, and Y. Itoh. A noise reduction method based on linear prediction with variable step-size. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E88-A(4):855–861, 2005.

Doh-Suk Kim and M. Tarraf. Enhanced perceptual model for non-intrusive speech quality assessment. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, 2006. doi: 10.1109/ICASSP.2006.1660149.

J. Kim and C. Un. Enhancement of noisy speech by forward/backward adaptive digital filtering. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 11, pages 89–92, April 1986.

Jong Uk Kim, S. G. Kim, and C. D. Yoo. The incorporation of masking threshold to subspace speech enhancement. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 76–79, April 2003. doi: 10.1109/ICASSP.2003.1198720.

Wooil Kim and R. M. Stern. Band-independent mask estimation for missing-feature reconstruction in the presence of unknown background noise. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 305–308, May 2006. doi: 10.1109/ICASSP.2006.1660018.

N. Kitawaki and T. Yamada. Subjective and objective quality assessment for noise reduced speech. In *ETSI Workshop on Speech and Noise in Wideband Communication*, Sophia Antipolis, France, May 2007.

E. J. Kreul, J. C. Nixon, K. D. Kryter, D. W. Bell, J. S. Lang, and E. D. Schubert. A proposed clinical test of speech discrimination. *J. Speech. Hear. Res.*, 11:536–553, 1968.

T. Kristjansson and J. Hershey. High resolution signal reconstruction. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 291–296, December 2003. doi: 10.1109/ASRU.2003.1318456.

T. Kristjansson, B. Frey, L. Deng, and A. Acero. Towards non-stationary model-based noise adaptation for large vocabulary speech recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 337–340, May 2001. doi: 10.1109/ICASSP.2001.940836.

D. A. Krubsack and R. J. Niederjohn. An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech. *IEEE Trans. Signal Process.*, 39(2):319–329, February 1991. doi: 10.1109/78.80814.

David A. Krubsack and Russell J. Niederjohn. Estimation of noise corrupting speech using extracted speech parameters and averaging of logarithmic modified periodograms. *Digital Signal Processing*, 4 (3):154–172, July 1994.

K. Kryter. Methods for the calculation and use of the articulation index. *J. Acoust. Soc. Am.*, 34(11): 1689–1697, 1962.

H. Law and R. Seymour. A reference distortion system using modulated noise. *Proc. IEEE*, 109B(48): 484–485, 1962.

K. Lebart, J. M. Boucher, and P. N. Denbigh. A new method based on spectral subtraction for speech dereverberation. *Acta Acoustica*, 87:359–366, 2001.

Te-Won Lee and Kaisheng Yao. Speech enhancement by perceptual filter with sequential noise parameter estimation. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 693–696, May 2004. doi: 10.1109/ICASSP.2004.1326080.

I. Lehiste and G. E. Peterson. Linguistic considerations in the study of speech intelligibility. *J. Acoust. Soc. Am.*, 31(3):280–286, 1959.

R. Leonard. A database for speaker-independent digit recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 328 – 331, March 1984.

H. Lev-Ari and Y. Ephraim. Extension of the signal subspace speech enhancement approach to colored noise. *IEEE Signal Process. Lett.*, 10(4):104–106, April 2003. doi: 10.1109/LSP.2003.808544.

H. Levitt. Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.*, 49(2):467–477, 1971.

Peng Li, Yong Guan, Bo Xu, and Wenju Liu. Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech. In *Proc. Intl Conf on Innovative Computing, Information and Control*, volume 2, pages 742–745, August 2006. doi: 10.1109/ICICIC. 2006.311.

J. S. Lim and A. V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proc. IEEE*, 67(12):1586–1604, December 1979.

Jae Lim. Speech enhancement. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume II, pages 3135–3142, April 1986.

L. Lin, E. Ambikairajah, and W. H. Holmes. Speech enhancement for nonstationary noise environment. In *Proc. Asia-Pacific Conf. on Circuits and Systems*, volume 1, pages 177–180, 2002a. doi: 10.1109/ APCCAS.2002.1114931.

L. Lin, W. H. Holmes, and E. Ambikairajah. Speech denoising using perceptual modification of Wiener filtering. *IEE Electronics Lett.*, 38(23):1486–1487, 2002b. doi: 10.10491el:20020965.

L. Lin, W. H. Holmes, and E. Ambikairajah. Adaptive noise estimation algorithm for speech enhancement. *IEE Electronics Lett.*, 39(9):754–755, 2003a. doi: 10.1049/el:20030480.

L. Lin, W. H. Holmes, and E. Ambikairajah. Subband noise estimation for speech enhancement using a perceptual Wiener filter. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 80–83, 2003b. doi: 10.1109/ICASSP.2003.1198721.

Z. Lin and R. Goubran. Instant noise estimation using Fourier transform of AMDF and variable start minima search. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, 2005. doi: 10.1109/ICASSP.2005.1415075.

Zhong Lin, Rafik A. Goubran, and Richard M. Dansereau. Noise estimation using speech/non-speech frame decision and subband spectral tracking. *Speech Communication*, 49(7–8):542–557, July 2007. doi: 10.1016/j.specom.2006.10.002.

W. M. Liu, K. A. Jellyman, J. S. D. Mason, and N. W. D. Evans. Assessment of objective quality measures for speech intelligibility estimation. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume I, pages 1225–1228, 2006. doi: 10.1109/ICASSP.2006.1660248.

P. Lockwood and J. Boudy. Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust recognition in cars. *Speech Communication*, 11:215–228, June 1992.

P. C. Loizou. *Speech Enhancement Theory and Practice*. Taylor & Francis, 2007. ISBN 978–0849350320.

T. Lotter, C. Benien, and P. Vary. Multichannel speech enhancement using Bayesian spectral amplitude estimation. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, April 2003.

Cedar Audio Ltd. Debuzz: Eliminating buzzes and hums. Forensic News 3, Cedar Audio Ltd, August 2005.

Cedar Audio Ltd. Adaptive filters. Forensic News 5,6,7, Cedar Audio Ltd, 2006.

C. Ludvigsen. Relations among some psychoacoustic parameters in normal and cochlearly impaired listeners. *J. Acoust. Soc. Am.*, 78:1271–1280, 1985.

M-Audio. Izotope RX - complete audio restoration software, 2007. URL `http://www.m-audio.com/products/en\_us/iZotopeRX-main.html`.

N. Magotra, F. Livingston, and S. Rajagopalan. Single channel speech enhancement in real time. In *Proc. Asilomar Conf. on Signals, Systems and Computers*, volume 2, pages 1211–1215, 1993. doi: 10.1109/ACSSC.1993.342379.

D. Malah, R. V. Cox, and A. J. Accardi. Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 789–792, March 1999. doi: 10.1109/ICASSP.1999.759789.

Kotta Manohar and Preeti Rao. Speech enhancement in nonstationary noise environments using noise properties. *Speech Communication*, 48(1):96–109, January 2006. doi: 10.1016/j.specom.2005.08.002.

R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process.*, 9:504–512, July 2001. doi: 10.1109/89.928915.

R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech Audio Process.*, 13(5):845–856, September 2005. doi: 10.1109/TSA.2005.851927.

R. Martin and R. V. Cox. New speech enhancement techniques for low bit rate speech coding. In *Proc. IEEE Workshop on Speech Coding*, pages 165–167, June 1999. doi: 10.1109/SCFT.1999.781519.

Rainer Martin. Spectral subtraction based on minimum statistics. In *Proc. European Signal Processing Conf*, pages 1182–1185, 1994.

M. Marzinzik and B. Kollmeier. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *IEEE Trans. Speech Audio Process.*, 10(2):109–118, February 2002. doi: 10.1109/89.985548.

K. Mayyas and T. Aboulnasr. Leaky LMS algorithm: MSE analysis for Gaussian data. *IEEE Trans. Signal Process.*, 45(4):927–934, 1997. ISSN 1053–587X. doi: 10.1109/78.564181.

R. McAulay and M. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust., Speech, Signal Process.*, 28(2):137–145, April 1980.

R. McAulay and T. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust., Speech, Signal Process.*, 34(4):744–754, August 1986.

B. McDermott. Multidimensional analysis of circuit quality judgements. *J. Acoust. Soc. Am.*, 45(3): 774–781, 1968.

G. Miller and P. Nicely. An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.*, 27(2):338–352, 1955.

U. Mittal and N. Phamdo. Signal/noise KLT based approach for enhancing speech degraded by colored noise. *IEEE Trans. Speech Audio Process.*, 8(2):159–167, March 2000. doi: 10.1109/89.824700.

D. Morgan, C. Kamm, and T. Velde. Form equivalence of the speech perception in noise (SPIN) test. *J. Acoust. Soc. Am.*, 69(6):1791–1798, 1981.

J. Mourjopoulos, P. Clarkson, and J. Hammond. A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 7, pages 1858–1861, May 1982.

J. N. Mourjopoulos. Digital equalization of room acoustics. *Journal Audio Eng. Soc.*, 42(11):884–900, November 1994.

W. Munson and J. Karlin. Isopreference method for evaluating speech transmission circuits. *J. Acoust. Soc. Am.*, 34(6):762–774, 1962.

T. Nakatani, M. Miyoshi, and K. Kinoshita. Single-microphone blind dereverberation. In J. Benesty, S. Makino, and J. Chen, editors, *Speech Enhancement*. Springer Verlag, April 2005.

S. T. Neely and J. B. Allen. Invertibility of a room impulse response. *J. Acoust. Soc. Am.*, 66(1):165–169, July 1979.

E. Nemer, R. Goubran, and S. Mahmoud. SNR estimation of speech signals using subbands and fourth-orderstatistics. *IEEE Signal Process. Lett.*, 6(7):171–174, 1999. doi: 10.1109/97.769361.

M. Nilsson, S. Soli, and J. Sullivan. Development of hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *J. Acoust. Soc. Am.*, 95(2):1085–1099, 1994.

NIST. NIST speech tools, 1992. URL `http://www.nist.gov/speech/tools`.

Noisevault. Impulse resonse repository, 2007. URL `http://noisevault.com/nv/`.

A. V. Oppenheim and R. W. Schafer. *Digital Signal Processing*. Prentice Hall, 1975.

A. V. Oppenheim, R. W. Schafer, and T. G. Stockham, Jr. Nonlinear filtering of multiplied and convolved signals. *IEEE Trans. Audio Electroacoust.*, AU-16(3):437–466, September 1968.

F. S. Pacheco and R. Seara. A single-microphone approach for speech signal dereverberation. In *Proc. European Signal Processing Conf. (EUSIPCO)*, Antalya, Turkey, September 2005.

B. L. Pellom and J. H. L. Hansen. An improved (Auto:I, LSP:T) constrained iterative speech enhancement for colored noise environments. *IEEE Trans. Speech Audio Process.*, 6(6):573–579, November 1998. doi: 10.1109/89.725324.

R. Plomp and A. Mimpen. Speech-reception threshold for sentences as a function of age and noise level. *J. Acoust. Soc. Am.*, 66(5):1333–1342, 1979.

Schuyler R. Quackenbush, Thomas P. Barnwell, III, and Mark A. Clements. *Objective Measures of Speech Quality*. Prentice Hall, January 1988. ISBN 0136290566.

Bhiksha Raj, Michael L. Seltzer, and Richard M. Stern. Reconstruction of missing features for robust speech recognition. *Speech Communication*, 43:275–296, 2004.

Karthikesh Raju. Parallel model combination - a primer, robustness in language and speech processing. Technical Report T-61.182, Helsinki Univ of Tech, June 2003.

S. Rangachari and P. C. Loizou. A noise-estimation algorithm for highly non-stationary environments. *Speech Communication*, 48(2):220–231, February 2006. doi: 10.1016/j.specom.2005.08.005.

S. Rangachari, P. C. Loizou, and Y. Hu. A noise estimation algorithm with rapid adaptation for highly nonstationary environments. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 305–308, May 2004. doi: 10.1109/ICASSP.2004.1325983.

S. Rennie, T. Kristjansson, P. Olsen, and R. Gopinath. Dynamic noise adaptation. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, May 2006. doi: 10.1109/ICASSP.2006.1660241.

A. Rezayee and S. Gazor. An adaptive KLT approach for speech enhancement. *IEEE Trans. Speech Audio Process.*, 9(2):87–95, 2001. doi: 10.1109/89.902276.

K. S. Rhebergen and N. J. Versfeld. A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *J. Acoust. Soc. Am.*, 117(4):2181–2192, 2005.

K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler. Release from informational masking by time reversal of native and non-native interfering speech. *J. Acoust. Soc. Am.*, 118(3):1274–1277, 2006.

Christophe Ris and Stephane Dupont. Assessing local noise level estimation methods: Application to noise robust ASR. *Speech Communication*, 34(1–2):141–158, April 2001. doi: 10.1016/S0167--6393(00)00051--0.

A. Rix. Perceptual speech quality assessment - a review. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, pages 1056–1059, 2004.

A. Rix, J. Beerends, M. Hollier, and A. Hekstra. Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 749–752, 2001.

J. Roberts. Modification to piecewise LPC-10E. Technical Report WP-21752, MITRE, 1978.

E. H. Rothauser, W. D. Chapman, N. Guttman, M. H. L. Hecker, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock. IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust.*, 17(3):225–246, 1969.

M. R. Sambur. Adaptive noise canceling for speech signals. *IEEE Trans. Acoust., Speech, Signal Process.*, 26(5):419–423, October 1978.

N. Sasaoka, K. Sumi, Y. Itoh, and K. Fujii. A new noise reduction system based on ALE and noise reconstruction filter. In *Proc. Intl. Symp. on Circuits and Systems*, pages 272–275 Vol. 1, 2005. doi: 10.1109/ISCAS.2005.1464577.

N. Sasaoka, M. Watanabe, Y. Itoh, and K. Fujii. Noise reduction system based on LPEF and system identification with variable step size. In *Proc. Intl. Symp. on Circuits and Systems*, pages 2311–2314, 2007. doi: 10.1109/ISCAS.2007.378850.

M. R. Schroeder, B. S. Atal, and J. L. Hall. Optimizing speech coders by exploiting masking properties of the human ear. *J. Acoust. Soc. Am.*, 66(6):1647–1652, 1979.

M. Seltzer, B. Raj, and R. Stern. A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition. *Speech Communication*, 43:379–393, September 2004.

Xuemin Shen and Li Deng. A dynamic system approach to speech enhancement using the h_∞ filtering algorithm. *IEEE Trans. Speech Audio Process.*, 7(4):391–399, July 1999. doi: 10.1109/89.771261.

J. Shore. Minimum cross-entropy spectral analysis. *IEEE Trans. Acoust., Speech, Signal Process.*, 29(2): 230–237, April 1981.

Jongseo Sohn and Wonyong Sung. A voice activity detector employing soft decision based noise spectrum adaptation. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 365–368, May 1998. doi: 10.1109/ICASSP.1998.674443.

Myung-Suk Song, Chang-Heon Lee, and Hong-Goo Kang. Performance analysis of various single channel speech enhancement algorithms for automatic speech recognition. In *Proc. Intl. Conf. on Spoken Lang. Processing (ICSLP)*, 2006.

V. Stahl, A. Fischer, and R. Bippus. Quantile based noise estimation for spectral subtraction and Wiener filtering. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, pages 1875–1878, 2000. doi: 10.1109/ICASSP.2000.862122.

G. B. Stan, J. J. Embrechts, and D. Archambeau. Comparison of different impulse response measurement techniques. *Journal Audio Eng. Soc.*, 50(4):249–262, 2002.

H. J. M. Steeneken and F. W. M. Geurtsen. Description of the RSG.10 noise data-base. Technical Report IZF 1988–3, TNO Institute for perception, 1988.

H. J. M. Steeneken and T. Houtgast. A physical method for measuring speech-transmission quality. *J. Acoust. Soc. Am.*, 67(1):318–326, January 1980.

H. J. M. Steeneken and T. Houtgast. Mutual dependence of the octave-band weights in predicting speech intelligibility. *Speech Communication*, 28:109–123, 1999.

Amarnag Subramanya, Zhengyou Zhang, Zicheng Liu, and Alex Acero. Speech modeling with magnitude-normalized complex spectra and its application to multisensory speech enhancement. Technical Report MSR-TR-2005–126, Microsoft Reseach, 2005.

A. Sugiyama, H. P. Hua, M. Kato, and M. Serizawa. Noise suppression with synthesis windowing and pseudo noise injection. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 545–548, 2002. doi: 10.1109/ICASSP.2002.1005797.

B. Frey T. Kristjansson and L. Deng. Joint estimation of noise and channel distortion in a generalized EM framework. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Madonna di Campiglio, December 2001.

D. Talkin. A robust algorithm for pitch tracking (RAPT). In W. B. Kleijn and K. K. Paliwal, editors, *Speech Coding and Synthesis*, pages 495–518. Elsevier, Amsterdam, 1995.

T. W. Tillman, P. C. Bucy, and R. Carhart. Monaural vs binaural discrimination for filtered CNC materials. In *The impaired auditory mechanism*. USAF school of aerospace medicine, Brooks Airforce Base, Texas, 1966.

A. P. Varga and R. K. Moore. Hidden Markov model decomposition of speech and noise. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 845–848, April 1990. doi: 10.1109/ICASSP.1990.115970.

A. P. Varga, H. J. M Steeneken, M. Tomlinson, and D. Jones. The NOISEX-92 study on the effect of additlve noise on automatic speech recognition. Technical report, DRA Speech Res. Unit, 1992.

Andrew Varga and Herman J. M. Steeneken. Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 3(3):247–251, July 1993. doi: 10.1016/0167--6393(93)90095--3.

L. Varner, T. Miller, and T. Eger. A simple adaptive filtering technique for speech enhancement. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 8, pages 1126–1128, April 1983.

Y. Vazquez-Alvarez and M. Huckvale. The reliability of the ITU-P.85 standard for the evaluation of text-to-speech systems. In *Proc. Intl. Conf. on Spoken Lang. Processing (ICSLP)*, pages 329–332, 2002.

Nathalie Virag. Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Trans. Speech Audio Process.*, 7(2):126–137, March 1999. doi: 10.1109/89.748118.

W. Voiers. Diagnostic acceptability measure for speech communication systems. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 204–207, 1977.

W. D. Voiers. Evaluating processed speech using the diagnostic rhyme test. *Speech Technology*, 1(4): 30–39, 1983.

E. Wan, A. Nelson, and Rick Peterson. Speech enhancement assessment resource (SpEAR) database. Database Beta Release v1.0, CSLU, Oregon Graduate Institute of Science and Technology, 1998. URL `http://cslu.cse.ogi.edu/nsel/data/whitePaper.html`.

D. Wang and Jae Lim. The unimportance of phase in speech enhancement. *IEEE Trans. Acoust., Speech, Signal Process.*, 30(4):679–681, 1982. ISSN 0096–3518.

DeLiang Wang and Guy Brown, editors. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley, 2006. ISBN 978–0471741091. URL `http://www.casabook.org`.

Waves. IR parametric convolution plugin. Technical report, 2007. URL `http://www.acoustics.net/`.

B. Widrow, J. R. Glover, Jr, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, Eugene Dong, Jr, and R. C. Goodlin. Adaptive noise cancelling: Principles and applications. *Proc. IEEE*, 63(12):1692–1716, 1975. ISSN 0018–9219.

K. Worrall, R. Fellows, J. Causer, and L. Craigie. Inteligibility testing at HM Government Communications Centre. *Proc. Institute of Acoustics*, 28(6):12, 2006.

M. Wu and D. Wang. A two-stage algorithm for one-microphone reverberant speech enhancement. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(3):774–784, May 2006.

T. Yamada, M. Kumakura, and N. Kitawaki. Word intelligibility estimation of noise-reduced speech. In *Proc. Interspeech*, pages 169–172, Pittsburgh, Pennsylvania, 2006.

J. Yang. Frequency domain noise suppression approaches in mobile telephone systems. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 363–366, April 1993. doi: 10.1109/ICASSP.1993.319313.

K. Yao and S. Nakamura. Sequential noise compensation by sequential Monte Carlo method. In *Advances in Neural Information Processing Systems*, volume 14, pages 1213–1220, 2002.

Kaisheng Yao and Te-Won Lee. Speech enhancement with noise parameter estimated by a sequential Monte Carlo method. In *IEEE Workshop on Statistical Signal Processing*, pages 609–612, October 2003. doi: 10.1109/SSP.2003.1289553.

A. Yasmin, P. Fieguth, and Li Deng. Speech enhancement using voice source models. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 797–800, March 1999. doi: 10.1109/ICASSP.1999.759791.

B. Yegnanarayana and P. S. Murthy. Enhancement of reverberant speech using LP residual signal. *IEEE Trans. Speech Audio Process.*, 8(3):267–281, May 2000.

Wei Zhang and S. Gazor. Statistical modelling of speech signals. In *Proc. Intl. Conf. on Signal Processing*, volume 1, pages 480–483, August 2002. doi: 10.1109/ICOSP.2002.1181096.