# Chapter 8: Acoustic Features and Distance Measure to Reduce Vulnerability of ASR Performance Due to the Presence of a Communication Channel and/or Background Noise

Johan de Veth, Bert Cranen & Louis Boves  $A^2RT$ , Department of Language and Speech University of Nijmegen, The Netherlands

# 1 Introduction

# 1.1 Automatic speech recognition is pattern recognition

Saying that late 20th century automatic speech recognition (ASR) is pattern recognition, is something of a truism, but perhaps one of which the fundamental implications are not always fully appreciated. Essentially, a pattern recognition task boils down to measuring the distance between a physical representation of a new, as yet unknown token, and all elements of a set of pre-existing patterns, of course in the same physical representation. On the one hand, the 'patterns' that can be recognized are, implicitly or explicitly, separate and invariable entities. For example, the command open in a Windows control application always has the same invariable and unique meaning. On the other hand, the unknown input tokens are continuous signals that typically show a high degree of variability. ASR research has centered around the problem of how to map continuous, variable acoustic representations onto discrete, invariable patterns. In ASR the physical representation of the speech tokens is some kind of dynamic power spectrum, for reasons which date back to the days of Ohm and von Helmholtz, who have shown that the power spectrum explains most of the perceptual phenomena in human speech processing. Since the inception of digital signal processing dynamic spectra are approximated by a sequence of short-time spectra (Rabiner and Schafer, 1978). Consequently, the pattern match in ASR is invariably implemented as the accumulation of some distance measure between the acoustic features derived from a sequence of short-time spectra of the input token and the corresponding representation of the active patterns (see Fig. 1). Therefore, anything which adds to the variability of the short time spectrum of a speech signal will, as it were by definition, complicate pattern matching, and consequently complicate ASR.

The basic aim of robust speech recognition is to make the pattern match insensitive to variability in the short-time spectra. It goes without saying that there is no single



**Figure 1.** Automatic speech recognition as a form of pattern matching. The feature extraction module is used to compute a compact representation of the short-time spectra that describes characteristics which are best suited for recognition.

optimal approach to find the *holy grail*. There are too many factors which affect the spectra; in addition, there are just too many different ways in which the spectral features can be represented, in which the patterns can be cast, and in which the search for the best matching pattern can be implemented, even if we restrict the discussion to hidden Markov models (HMMs) and Viterbi search. In this contribution we investigate two factors which increase the variability of the short-time spectra in more detail, viz. the transmission channel and background noise. To clarify the discussion we first sketch a model of the speech signal at the input of an ASR device. With the help of this model we can explain the relation between different approaches of robust speech recognition, and show how these approaches can be combined to reduce the effects of different sources that distort the short-time spectra.

# **1.2** A simple model

It is easy to draw a comprehensive, but very abstract, conceptual model of the signals at the input of an ASR device. Such a model is depicted in Fig. 2. We always have a speaker, who is in some physical, acoustic environment, talking into a microphone which is connected to a transmission chain that eventually delivers a signal to the input of the ASR device. Filling in the details of this abstract model is less easy and straightforward, however. How must one model the acoustic environment of the speaker? The model of an anechoic room is certainly different from the model of a car running on the highway, and also different from the model of a busy train station, and from a quiet hotel room. On top of the impact of the acoustic environment, the microphone that converts the sound into an electrical signal has several effects of its own. It is not equally sensitive to all frequencies, nor to sounds coming from all different directions. In addition, the microphone may also introduce non-linear distortions, like the plops caused by the airflow in fricative and/or plosive sounds. In short, the microphone may introduce linear and non-linear effects. The transmission chain connecting the microphone to the ASR device



**Figure 2.** The observed speech spectrum S' is a mix of contributions from the original speech spectrum S, the background noise spectrum N, the linear and non-linear transfer characteristics of the channel, which can be described by a series of Volterra kernels  $H_1, H_2, \ldots$  (Schetzen, 1980), and the spontaneous activity of the channel  $H_0$ .

may introduce its own additive noise and linear filtering (e.g., caused by the cabling and amplifiers in analogue telephone transmission), as well as non-linear distortions. Digital transmission should help to alleviate channel distortions, but it is certainly no panacea. In cellular digital telephone networks radio transmission errors make a very substantial contribution to recognition errors (both human and machine, for that matter).

Of course, we build conceptual models of the signal at the input of the ASR device with the aim to recover the original undistorted speech signal, or rather, to recover the power spectrum of the undistorted signals. The mathematical expression describing the relation between the clean speech spectrum and the spectrum at the ASR input is dependent on (1) the details of the transmission chain that are accounted for, and (2) the description of the linear and non-linear effects that were mentioned above.

For the minimalist model shown in Fig. 2 let S' denote the spectrum of the speech signal at the ASR input. Clearly, S' is a function of time t and frequency  $\omega$ :  $S' = S'(t, \omega)$ . If we assume that (1) the background noise is additive and (2) the channel can be described as a linear, finite memory system,  $S'(t, \omega)$  can be written as

$$S'(t,\omega) = H_0(t,\omega) + H_1(t,\omega)[S(t,\omega) + N(t,\omega)],$$
(1)

where  $S(t, \omega)$  is the original speech spectrum and  $N(t, \omega)$  the spectrum of the background noise.  $H_0(t, \omega)$  corresponds to the spectrum of the signal spontaneously produced by the channel and  $H_1(t, \omega)$  is the spectrum of the linear transfer function of the channel (Schetzen, 1980). S', S, N,  $H_0$  and  $H_1$  are complex-valued functions of time and frequency.

The original 'clean' speech spectrum  $S(t, \omega)$  is very much an abstract concept. Even under quiet, 'noise free' conditions the clean signal cannot be observed, because it is affected by the room acoustics and the recording equipment. Fortunately, the lion's share of the inevitable effects are very small, much smaller than the differences between two speakers or between two realizations of the same linguistic utterance by the same speaker. As long as the effects are similar for all utterances, they can be considered as part of the 'clean' spectrum  $S(t, \omega)$ . Robust speech recognition comes into play where the impact on the 'clean' speech is variable, and so strong that the contributions to  $S'(t, \omega)$  can no longer be neglected.

For robust speech recognition, Eq. 1 can be used to tell us how the speech spectrum observed at the ASR input  $S'(t, \omega)$  can be understood in terms of the original speech spectrum  $S(t, \omega)$  on the one hand, and the distortion terms on the other, i.e., the spontaneous activity of the transmission channel  $H_0(t, \omega)$ , the linear transfer function of the channel  $H_1(t, \omega)$ , and the power spectrum of the additive noise  $N(t, \omega)$ . Any approach to robustness will need to consider (1) the relative importance of each of the three distortion terms  $H_0(t,\omega)$ ,  $H_1(t,\omega)$ , and  $N(t,\omega)$ , and (2) the accuracy with which each term can be estimated. We are now in the position to explain why under certain conditions specific solutions are superior. The key issue here is the amount of available prior knowledge. To illustrate our point, we consider the following two scenarios. First, take an ASR device attached to a switch in a telephone network. In this case, things look pretty hopeless. The input is a single signal, from which the set of actual parameter values of all components of the model in Fig. 2 must be estimated. From a mathematical point of view this is an ill-posed problem. It is an attempt to find a unique solution based on one equation with many unknowns. Elementary algebra tells us that this is impossible. Instead of a unique solution, a whole family of solutions is possible and without additional and independent observations the 'true' parameter values cannot be reconstructed. Consequently, in this situation we are obliged to simplify the model as much as possible (to reduce the number of unknowns). As we will see below, the well-known cepstrum mean subtraction technique (Atal, 1974; Furui, 1981) is a typical example of this simplified approach. For a second scenario, consider an ASR device built into a Bugatti car that is almost exclusively driven on the highways of Arizona and New Mexico. In this case, we might be able to reliably estimate the parameters of most components of the model in Fig. 2. To be more specific, with the engine switched off and no speech input we can establish the parameter values for the model component representing the spontaneous activity of the channel. While driving the car and no speech input, we can record the typical background noise. Finally, we can measure the linear transfer characteristic of the channel by testing the microphone in an an-echoic room. Now, the most appropriate robustness approach in the first scenario will seem much more primitive than the approach in the Bugatti case. Although it may be considered more appropriate from a 'physical' point of view, the more sophisticated Bugatti model would fail miserably in the first scenario, because the parameters of the noise will vary substantially between calls, making it impossible to come up with a useful prediction for an individual call. It is common knowledge in the field of System Identification that it is better to have an overly simplistic model of which the parameters can be estimated reliably, than to try a physically more adequate model, the parameters of which cannot be estimated reliably.

In the following Section, we will discuss techniques for dealing with linear filtering effects caused by the microphone and transmission channel. In Section 3, we will discuss different methods to deal with additive noise. In both problems, we will stress the inter-dependence between the underlying models, the choice for parameter representations, and the eventual spectral distance computation which is at the heart of any ASR algorithm.

# 2 The presence of a transmission channel

# 2.1 Assumptions for channel robustness

As said before, it is only possible to observe the speech signal through some transmission channel. Under most practical circumstances it is reasonable to assume that the linear transfer function of the channel is time-invariant or at least varying slowly in comparison with the articulation process. Then  $H_1(t, \omega)$  in Eq. 1 reduces to

$$H_1(t,\omega) = H_1(\omega). \tag{2}$$

In this Section, we want to focus the discussion on techniques for dealing with the effects of  $H_1(\omega)$ . To simplify the discussion, we make the additional assumption that the energy of the combination of the spontaneous activity of the channel and the background noise can be neglected, i.e.,

$$|H_0(t,\omega) + H_1(\omega)N(t,\omega)|^2 \ll |H_1(\omega)S(t,\omega)|^2.$$
 (3)

The speech spectrum at the ASR input can then be approximated as

$$S'(t,\omega) = H_1(\omega)S(t,\omega).$$
(4)

As can be seen, the speech spectrum at the ASR input now contains only two contributions: the time-invariant linear transfer function of the channel and the time-variant speech spectrum. As we will see below, the separation into two contributions that have different temporal characteristics is the key to many channel robustness strategies. Without the assumption expressed by Eq. 2 such a convenient separation is not possible.

In the log-energy domain Eq. 4 becomes

$$\log(|S'(t,\omega)|^2) = \log(|H_1(\omega)|^2) + \log(|S(t,\omega)|^2).$$
(5)

Taking the Fourier transform, we have in the cepstral domain

$$c'(t,\tau) = c_h(\tau) + c(t,\tau),$$
(6)

with  $c'(t,\tau)$  the cepstrum of the channel output,  $c_h(\tau)$  the cepstrum of the channel and  $c(t,\tau)$  the cepstrum of the original speech signal.  $c'(t,\tau)$ ,  $c_h(\tau)$ , and  $c(t,\tau)$  are realvalued functions, because the terms in Eq. 5 ( $\log(|S'(t,\omega)|^2)$ , etc.) are real and even. According to Eq. 6 the contribution of the unknown channel is a constant for a given quefrency  $\tau$ . From Eqs. 4, 5 and 6 it can be understood how the channel can affect ASR: If training and testing are performed using two different channels, the cepstra used during training differ from the corresponding cepstra at recognition time. Channel normalization methods aim to reduce the differences between training and testing speech spectra caused by the channel.

# 2.2 Channel normalization techniques

Channel normalization (CN) techniques have been studied for quite different conditions. In one such condition, which is not addressed in this contribution, a recognizer is trained with speech recorded with a close talking microphone and recognition is attempted on speech recorded with a different microphone. The contribution of the channel during training does not need to be known in great detail, because it is constant. The channel conditions during test are different from those during training, but constant too. Thus, a single, fixed transformation should suffice (see for example (Anastasakos et al., 1994; Liu et al., 1994; Orloff et al., 1994; Weintraub et al., 1994)).

In ASR applications over the telephone the situation is different: the channels over which the training speech is recorded are unknown and likely to differ between all recording sessions. The same goes for the testing speech. Under these conditions a CN technique is needed both during training and testing. This is the situation that is reviewed here.

#### 2.2.1 Use of many different channels

Under the assumption that the channel characteristics and the speech signal are statistically independent, the first and second order statistics of the cepstral parameters for a particular pattern (in our case a subword unit s) can be written according to Eq. 6 as

$$E\{c'_{s}(t,\tau)\} = E\{c_{h}(\tau)\} + E\{c_{s}(t,\tau)\}$$
(7)

and

$$Cov\{c'_{s}(t,\tau)\} = Cov\{c_{h}(\tau)\} + Cov\{c_{s}(t,\tau)\},$$
(8)

where  $E\{\}$  denotes the expected value and  $Cov\{\}$  denotes the covariance. Thus, the emission probability density functions of the states associated with s contain a contribution due to the statistical differences between the channels. The estimates of the means are biased with the unknown function  $E\{c_h(\tau)\}$ , which approaches the average channel cepstrum if enough different channels are used. At the same time, the covariance estimates are increased by the factor  $Cov\{c_h(\tau)\}$ . If the training speech covers a sufficient number of different but representative channels, the estimated parameters of the emission probability density functions may be expected to adhere to Eqs. 7 and 8, e.g., (Hermansky et al., 1991; Hirsch et al., 1991; Aikawa et al., 1993; Haeb-Umbach et al., 1995; Junqua et al., 1995; Nadeu et al., 1995a; Singer et al., 1995; de Veth and Boves, 1996). Thus, using many different channels in training helps to reduce the impact of specific channels on the eventual models. However, it is of limited help when an unknown utterance must be recognized, because there is no guarantee that the bias due to the particular channel at hand is close to the average channel  $E\{c_h(\tau)\}$ .

#### 2.2.2 Explicit channel estimation

Some authors have proposed to estimate the contributions of the channel explicitly (statistical channel modeling (Gish et al., 1985; Gish et al., 1986)). This estimate is then used to correct the HMM parameters or the speech feature values. The well-known cepstrum mean subtraction technique (Atal, 1974; Furui, 1981) can be considered as a form of explicit channel estimation. The cepstrum mean  $\hat{c}_{mean}(\tau)$  is computed by taking the average over all speech signal frames

$$\hat{c}_{mean}(\tau) = \frac{1}{T} \sum_{t=1}^{T} [c'(t,\tau) + c_h(\tau)] = \hat{c}_{utt}(\tau) + \hat{c}_h(\tau).$$
(9)

As indicated in Eq. 9, the cepstrum mean consists of the average speech cepstrum of the utterance  $\hat{c}_{utt}(\tau)$  (an approximation of the average speech cepstrum of the person who produced the utterance) and the contribution due to the channel  $\hat{c}_h(\tau)$ . As can be inferred from Eq. 9, cepstrum mean subtraction has two effects. Firstly, the variability that is due to differences between speakers is reduced in the statistics for a particular speech sound. Secondly, cepstrum mean subtraction reduces the variability due to differences between communication channels used when recording that sound.

Note that non-speech signal portions are not used in Eq. 9. The reason for this can be understood as follows. According to Eq. 1, the observed spectrum in non-speech signal portions is

$$S'(t,\omega) = H_0(t,\omega) + H_1(t,\omega)N(t,\omega).$$
<sup>(10)</sup>

As can be seen, the non-speech spectrum consists of two contributions: the spontaneous activity of the channel  $H_0(t, \omega)$  and the linearly filtered background noise  $H_1(\omega)N(t, \omega)$ . Without additional assumptions about the spontaneous activity of the channel  $H_0(t, \omega)$  and the background noise  $N(t, \omega)$ , non-speech portions of the signal cannot be used to obtain reliable information about the linear filter characteristic of the channel  $H_1(\omega)$  alone. As a consequence, using non-speech portions of the signal introduces bias in the estimate of the mean cepstrum in a way that cannot be predicted.

#### 2.2.3 Filtering of log-energy or cepstral feature values

It is well-known that any differentiable function f(t) can be recovered (up to a known constant) as follows when it is observed with a constant unknown bias k:

$$f(t) - f(t_{low}) = \int_{t_{low}}^{t} \frac{d}{dt'} [f(t') + k] dt',$$
(11)

where  $t_{low}$  satisfies  $-\infty < t_{low} < t$ . Due to the differentiate and re-integrate operation the unknown bias term k is replaced by the constant term  $f(t_{low})$ .

According to Eq. 6, the contribution of the channel results in a time-invariant additive bias for each cepstral coefficient, which is independent of the original speech cepstrum. When applying Eq. 11 to the cepstrum observed at the channel output, the

cepstrum will no longer depend on the channel after such a differentiate and re-integrate operation. Recalling our assumption in Eq. 3 about the energy of the silent signal portions being very small compared to the energy of speech portions, we see that the contribution  $f(t_{low})$  will vanish if we take  $t_{low}$  in a silent portion (e.g., in the leading silence at the beginning of the utterance). As a result, the cepstrum of the original speech signal is obtained.

The differentiate-and-integrate operation can be implemented as a linear filter. Filtering can be performed either in the log-energy domain (Hermansky et al., 1991; Hermansky and Morgan, 1994) or in the cepstral domain (Haeb-Umbach et al., 1995; de Veth and Bourlard, 1995; de Veth and Boves, 1996). It was shown that channel robustness can also be improved if the re-integration of Eq. 11 is omitted (Furui, 1981; Soong and Rosenberg, 1986). In that case, channel robustness is improved because the constant bias term k actually corresponds to DC in the modulation spectrum and the differentiation effectively attenuates this DC component.

If a properly designed leaky integrator is used, the differentiate-and-integrate operation will also be effective if the channel transfer function is slowly time-varying. This approach of the unknown channel problem has resulted in many different proposals for filtering the observed sequence of cepstral parameters, for instance RASTA filtering (Hermansky et al., 1991; Hermansky and Morgan, 1994), the Gaussian dynamic cepstrum representation (Aikawa et al., 1993; Singer et al., 1995; Boda et al., 1996), the high-pass filter method proposed in (Hirsch et al., 1991), Slepian filters (Nadeu et al., 1995a), phase-corrected RASTA (de Veth and Boves, 1996; de Veth and Boves, 1997b) and combinations of these methods (Junqua et al., 1995).

The cepstrum mean subtraction technique (Atal, 1974; Furui, 1981) can also be formulated as a linear filter operation. If the channel estimate is calculated over the full length of the speech utterance (as can be done in off-line experiments), then cepstrum mean subtraction can be interpreted as a FIR filter operation, with the filter adjusted to the length of each utterance. If a running mean is used to obtain the channel estimate (which is the common *modus operandi* for cepstrum mean subtraction in real-time applications) the definition of the FIR filter is the same for each utterance. The Gaussian Dynamic Cepstrum Representation (Aikawa et al., 1993) and the discrete cosine transform (Milner and Vaseghi, 1995) are other examples of FIR filtering, although none of these techniques was originally presented in that formalism.

# 2.3 Comparison of feature track filtering techniques

Using the linear transfer function description of Eq. 4, it is easy to show that the differences between the many techniques for undoing the effect of the channel relate to either the way in which the parameters of the linear filter  $H_1(\omega)$  are estimated, and/or to the way in which the operation, used to undo the effect of  $H_1(\omega)$  is implemented. In this Section, the focus is on the effects of the implementation of the filter. We will show that the details of the implementation are important, because these details interfere with the type of models that are used in speech recognition. Although it would have been interesting to show that the arguments developed in this Section also hold for more sophisticated models of the channel, and the attendant more complex techniques to undo these effects in a recogniser (Rahim and Juang, 1996; Junqua and Haton, 1996; Junqua, 2001), we will limit ourselves to the techniques decribed below due to space limitations.

# 2.3.1 Effect of the filter phase response

We concentrate our discussion on a comparison of three CN techniques which can be described in terms of cepstral filtering: RASTA filtering, cepstrum mean subtraction and phase-corrected RASTA. RASTA filtering (Hermansky et al., 1991; Hermansky and Morgan, 1994) is interesting for several reasons. First, it is well-known that RASTA filtering is effective with units that incorporate context dependency such as words or triphones. In addition, RASTA can be viewed as a crude model of auditory time masking, and it has been argued that this correspondence to perception accounts for much of its effectiveness (Hermansky and Pavel, 1995; Hermansky, 1996). However, Eq. 11 was the original point of departure for RASTA (Hermansky et al., 1991); its relation to auditory masking was only established later on (Hermansky and Pavel, 1995; Hermansky, 1996). The second CN technique discussed here is cepstrum mean subtraction (Atal, 1974; Furui, 1981), because this technique is very simple, yet highly effective (Steinbiss et al., 1995; Haeb-Umbach et al., 1995). In two independent studies the effectiveness of RASTA filtering and cepstrum mean subtraction was compared in a recognition set-up based on context independent HMMs (CI-HMMs) (Haeb-Umbach et al., 1995; de Veth and Boves, 1998a). In both studies the task was recognition of digit strings, be it that different languages were being used: (Haeb-Umbach et al., 1995) used German and American English, while (de Veth and Boves, 1998a) used Dutch. It was found in both studies that RASTA filtering is effective as a CN technique, but that cepstrum mean subtraction is more effective. The apparent limited effectiveness of RASTA filtering can be attributed to the wellknown left-context dependency introduced by the RASTA filtering (Koehler et al., 1994; Hermansky and Morgan, 1994). To be able to better understand this left-context dependency, and to be able to verify that this is indeed the underlying cause for the limited effectiveness of RASTA filtering when compared to cepstrum mean normalization, phasecorrected RASTA was introduced in (de Veth and Boves, 1998a).

We start our discussion by considering the signal shown in the upper panel of Fig. 3. This artificial signal is intended to represent a sequence of cepstral values for quefrency  $\tau$ . It models a sequence of seven time-invariant 'speech' states, preceded and followed by a rest state ('silence'). The signal contains a constant DC-component that represents the effect of the channel. The RASTA filtered version of this signal is shown in the middle panel of Fig. 3.

Two observations can be made. First, the DC-component has been removed (at least for times larger than, say, 70 frames). Second, the shape of the signal has been



**Figure 3.** Synthetic signal representing one of the cepstral coefficients in the feature vector. Upper panel: Original signal containing a time-invariant DC-offset. Middle panel: RASTA filtered signal. Lower panel: Phase-corrected RASTA filtered signal.

altered. Originally, the states of the signal had a constant amplitude. After filtering, the amplitude for each state drifts towards zero, while the values immediately after an abrupt change are more or less preserved. This explains why RASTA enhances the dynamic parts in the spectrum of a speech signal (Hermansky and Morgan, 1994). However, a description of the signal in terms of states with well-defined means and small variances becomes less accurate. Worse even, the mean amplitude of each state has become a function of the state itself AND of the preceding states. This is the left-context dependency in RASTA (Koehler et al., 1994; Hermansky and Morgan, 1994).

To identify the origin of this left-context dependency, we take a close look at the frequency response  $H_R(\nu)$  of the classical RASTA filter which can be written as

$$H_R(\nu) = |H_R(\nu)| e^{j\phi_R(\nu)},$$
(12)

with  $\nu$  the modulation frequency (in radians),  $|H_R(\nu)|$  the magnitude response and  $\phi_R(\nu)$  the phase response. The magnitude and phase response of the RASTA filter with integration factor a = -0.94 are shown in Fig. 4AB for modulation frequencies between 0-20 Hz. This range includes the 2 - 16 Hz region which has been shown to be most important for human speech recognition (Drullman et al., 1994). From Fig. 4B it can be seen that the phase response is non-linear for modulation frequencies below approximately 3 Hz. This non-linearity causes the time-domain shape distortions observed in the middle panel of Fig. 3.

To compensate for the phase distortion of the RASTA filter, while preserving its magnitude response, we followed a procedure suggested in (Hunt, 1978). After the



**Figure 4.** A. Log-energy response of classical RASTA. B. Phase response of classical RASTA. C. Log-energy response of phase-corrected RASTA. D. Phase response of phase-corrected RASTA.

RASTA filter an all-pass filter is applied whose phase response  $\phi_{pc}(\nu)$  is exactly the inverse of the phase response of the RASTA filter

$$\phi_{pc}(\nu) = -\phi_R(\nu). \tag{13}$$

Thus, the frequency response  $H_{pc}(\nu)$  of the phase correction filter is

$$H_{pc}(\nu) = e^{-j\phi_R(\nu)}.$$
 (14)

With this phase correction, the frequency response  $H_{pcR}(\nu)$  of the complete phasecorrected RASTA filter is

$$H_{pcR}(\nu) = H_R \times H_{pc} \qquad = |H_R(\nu)|. \tag{15}$$

The phase correction filter  $H_{pc}(\nu)$  can be implemented as a pole-zero filter, obtained by solving for the coefficients {b,a} that satisfy

$$e^{-j\phi_R(\nu)} = \frac{b_0 + b_1 e^{-j\nu} + \ldots + b_q e^{-jq\nu}}{1 + a_1 e^{-j\nu} + \ldots + a_p e^{-jp\nu}},$$
(16)

where q (p) is the order of the numerator (denominator) polynomial. In (de Veth and Boves, 1998a) a Matlab procedure with q = 1 and p = 7 was used to calculate the {b,a} coefficients (Little and Shure, 1993). As it turns out, three of the seven poles of the phase-correction filter are lying outside the unit circle, while the zero is lying inside. Due to the poles lying outside the unit circle, the phase-correction filter is unstable. Therefore, it cannot be applied directly to the RASTA filtered signal. However, the inverse of this filter is stable and, as a result, in off-line experiments the following engineering trick can be used (Hunt, 1978): (1) reverse the RASTA-filtered signal in time, (2) take the inverse of the pole-zero phase-correction filter, (3) apply the inverted phase-correction filter to the time-reversed RASTA-filtered signal and (4) reverse the resulting signal in time. With a slight performance penalty the non-causal filtering can be cast in a form that allows a real-time implementation with short processing delay (de Veth and Boves, 1997a). In this paper we will only discuss results for off-line experiments.

In Fig. 4CD the magnitude and phase response of the phase-corrected RASTA filter are shown. It can be seen that the magnitude response is almost identical to the original one and that the new phase response is flat and very close to zero in the region of important modulation frequencies. The result for phase-corrected RASTA in the time domain is shown in the lowest panel of Fig. 3. The shape of the phase-corrected RASTA filtered signal closely resembles that of the original signal. The phase correction effectively removes the amplitude drift towards zero in time-invariant parts of the signal and decreases the left-context dependency. Thus, phase-corrected RASTA does not model temporal masking, but it is in better agreement with the usual model of a speech utterance as a sequence of time-invariant states.

#### 2.3.2 Continuous speech recognition with phase-corrected RASTA

We compared the recognition performance of classical RASTA, phase-corrected RASTA and cepstrum mean subtraction for a continuous speech recognition task, where utterances recorded over the (land-line) public switched telephone network were used. About nine hours of speech were used for training, while three hours of speech were used for testing. The recognition lexicon contained 983 words. 1.2% of the words in the test set were out-of-vocabulary. The test set perplexity of the recognized sentence. Full details of these experiments can be found in (de Veth and Boves, 1998b).

We trained and tested HMMs for four different channel normalization conditions, i.e., no channel normalization (NCN), classical RASTA (clR), cepstrum mean subtraction (CMS) and phase-corrected RASTA (pcR) in combination with two different recognizer set-ups, i.e. context independent phone-based HMMs (CI-HMMs) and context dependent phone-based HMMs (CD-HMMs). In these experiments, the off-line versions of CMS and pcR were used. In other words, we used the whole utterance when we computed the cepstrum mean and when we applied the time-reversal operation needed for pcR. Taking the number of substitution, deletion and insertion errors into account, we computed the word error rate for all combinations of channel normalization method and recognizer set-up, where we varied the number of Gaussians used to describe the



**Figure 5.** Recognition accuracy as a function of the total number of Gaussians in the trained HMM set for clR ( $\times$ ), pcR ( $\star$ ) and CMS( $\bullet$ ), compared to the feature set without CN ( $\circ$ ) when using CI-HMMs.

emission probability density function of each state. For the CI-HMMs, mixtures with 4, 8, 16 and 32 Gaussians per state were used. This corresponds to using a total number of Gaussians of 460, 920, 1840 and 3680 respectively. The results for the different CN techniques with CI-HMMs are shown in Fig. 5.

From Fig. 5 it can be seen that clR deteriorates recognition performance compared to NCN, when CI-HMMs are used. Removing the channel bias by using clR introduces so much left-context dependency that the potential CN gain is completely annihilated. The results for pcR indicate that the poor performance of classical RASTA is a direct consequence of the phase distortion. By removing the phase distortion the recognition performance is significantly and substantially improved compared to clR. At the highest total number of Gaussians in this CI-HMM system the WER is reduced by 23% relative to clR. In addition, for the more complex acoustic models pcR recognition performance is significantly better than NCN and in fact becomes as good as CMS.

It is interesting to compare these results to the results reported in (de Veth and Boves, 1998a). Whereas the continuous speech results in Fig. 5 show that clR actually decreases recognition performance relative to NCN, the digit experiments in (Haeb-Umbach et al., 1995; de Veth and Boves, 1998a) showed that clR is viable as a CN technique. These findings may seem contradictory at first glance, but can be understood if one realizes that the main difference between these two set-ups is the number of different phone contexts. In fact, the number of different phone contexts for the continuous speech recognizer is more than 70 times as large as in the digit recognizer (de Veth and Boves, 1998b). As a consequence, the loss of recognition perfor-



**Figure 6.** Recognition accuracy as a function of the total number of Gaussians in the trained HMM set for clR ( $\times$ ), pcR ( $\star$ ) and CMS( $\bullet$ ), compared to the feature set without CN ( $\circ$ ) when using CD-HMMs.

mance due to enhancement of the left-context dependencies is likely to be much more important in the continuous speech recognizer. Apparently, this effect is so strong that it completely annihilates the potential performance gain obtained from the attenuation of modulation components near DC. In the case of the digit recognizer, the net effect of RASTA filtering is still positive, because the performance gain obtained by suppressing the DC component is less affected by the left-context effect. The number of different contexts for the digits vocabulary is apparently so small that the models are effectively context dependent.

When using clR with CD-HMMs one would expect that the loss of recognition performance due to the left-context effect is diminished, because different contexts are modeled with different states. When every individual left context could be modeled independently, the left-context effect should disappear completely, and the CN effect should remain in its full strength. However, under all practical conditions in continuous speech recognition, the amount of training data is not sufficient to model each left context independently. This lack of training data forces one to pool the data from different contexts for sub-word units with low occurrence counts. In our experiments the data sharing for infrequent units was implemented as a data-driven state-tying mechanism. Due to the data sharing, one can no longer expect that the loss of recognition performance caused by the left-context effect of clR is completely annihilated.

The results for CN techniques with CD-HMMs are shown in Fig. 6. In this set-up we used HMMs with 1, 2, 4 and 8 Gaussians per state, corresponding to a total number of Gaussians of 388, 776, 1552 and 3104 respectively. First, it can be seen

that the difference between clR and NCN has become smaller than the one we observed for CI-HMMs. For the best CI-HMMs clR decreases recognition performance by 16% relative to NCN. In the case of the best CD-HMMs the performance only drops 9%. Thus, we have some gain when switching from CI- to CD-HMMs in the case of clR, but this improvement is limited due to the state-tying mechanism that is used to avoid undertraining. However, even with CD-HMMs the detrimental effect of the left-context dependency is still stronger than the beneficial effect of CN in this recognition task.

It can also be seen that introducing the phase-correction brings the recognition performance curve very close to the one for CMS (except at the models corresponding to 1 Gaussian per state). For the CD-HMMs corresponding to 8 Gaussians per state, WER is improved by 15% when clR is replaced by pcR. This is in good agreement with the results of pcR obtained for CI-HMMs.

#### 2.4 Conclusions

Most, if not all techniques intended to eliminate the variability introduced by the communication channel imply some form of filtering of the sequence of feature vectors. The results of the comparison of three different techniques for CN and the explanation of these results show that care must be taken that the phase response of the filter is linear. In other words: The overall shape of the feature track must be preserved as much as possible. This requirement is due to the structure of the basic patterns which represent the spoken words. In most cases speech is modeled as a sequence of essentially timeinvariant states, which are only dependent on a very local context. Any phase distortion caused by a filter that is applied to remove the influence of the channel by necessity interferes with the time-invariance and independence assumptions. In the particular case of classical RASTA, the signal segments represented by conventional sub-word units are much shorter than the RASTA filter memory. This results in a conflict between the intrinsic nature of the feature values after filtering and the assumptions underlying the structure of the speech model. As a consequence, the intended beneficial effect of this CN technique is completely destroyed by the negative effect of the phase-distortion.

These findings show that any technique to improve robustness can only be expected to yield improved recognition results as long as it is compatible with the basic assumptions made in the models of the speech signal and in the procedure to search for the best matching patterns. Specifically, and in a way unfortunately, this implies that findings from human speech perception cannot simply be re-used in ASR algorithms which model speech as a sequence of discrete, time-invariant, context insensitive units.

#### **3** Robustness against background noise

#### **3.1** Assumptions for noise robustness

To understand the effect of acoustic background noise on the feature values at the input of an ASR system, we must return to Eq. 1. We keep the assumption that the channel characteristics are time-invariant (cf. Eq. 2), but we drop the additional assumption that the magnitudes of the terms related to noise are negligible relative to the magnitudes of the terms related to the speech signal. This is equivalent to the assumption that

$$|H_0(t,\omega) + H_1(\omega)N(t,\omega)| \approx |H_1(\omega)S(t,\omega)|.$$
(17)

We then have

$$S'(t,\omega) = H_0(t,\omega) + H_1(\omega)[S(t,\omega) + N(t,\omega)].$$
(18)

Without much loss of generality, Eq. 18 can be simplified by lumping all additive components into a single, possibly time-varying, noise component  $U(t, \omega)$ :

$$S'(t,\omega) = H_1(\omega)S(t,\omega) + U(t,\omega).$$
<sup>(19)</sup>

The general model in Eq. 19 is a good starting point for discussing different approaches for improving noise robustness. As we already discussed in Subsection 1.2, the choice for a particular approach depends on the assumptions that can be made about our knowledge of  $U(t, \omega)$ , or perhaps more accurately, on the possibility to obtain useful parametric estimates of  $U(t, \omega)$  in a specific situation. The model of  $U(t, \omega)$  must be more simple as the noise becomes more variable between situations and unpredictable for a specific situation. Attempts to undo the effect of additive noise can be classified according to their working domain. Popular approaches include methods (1) to clean the acoustic features, (2) to adapt the models trained on clean speech to noisy conditions and (3) to adapt the distance computation in the Dynamic Programming search. These approaches essentially try to reduce the variation in the feature values due to the noise or they try to limit the impact of this variation on the computation of the similarity between new observations and pre-existing models.

It is reasonable to assume that noise robustness of an ASR system will increase if methods developed in the different domains are properly combined. Although it might seem attractive to compare the performance of individual approaches for improved noise robustness, such a straightforward comparison is hardly fair. Some approaches may be inherently more effective with certain types of distortions. In addition, experience has shown that the effectiveness of robustness techniques may be dependent on the details of the implementation. For these reasons, we refrain from making direct comparisons between different noise robustness approaches. We will limit the discussion to summary descriptions of observation cleaning, predictive model-based compensation and model adaptation, and focus in more detail on a new approach, which is formulated in the local distance computation domain. For an extensive review of observation cleaning (and other noise robustness techniques), we refer to (Gong, 1995). Recently, many predictive model-based compensation schemes were reviewed in (Gales, 1998). Finally, we refer to (Lee, 1998) for an excellent review of model compensation and model adaptation techniques.

# **3.2** Three domains for noise robustness

#### 3.2.1 Feature domain

We first consider a well-known noise robustness method which is defined in the feature domain. In those scenarios where it is reasonable to assume that the noise is quasi time-invariant, an obvious strategy would be to make an estimate of the noise spectrum  $U(\omega)$  and to subtract it from the noisy input spectra  $S'(t, \omega)$ . This strategy is known as *spectral subtraction* (Boll, 1979; Lockwood and Boudy, 1992) and has a long tradition in research in speech enhancement, i.e., processing of noisy speech to make it more pleasant and intelligible for humans. Spectral subtraction can be regarded as a classic example of the idea to try and find a feature representation for which the statistical characteristics are minimally affected by the background noise. As long as ways can be found to reliably estimate the background noise characteristics, spectral subtraction is a useful pre-processing step that will increase recognition robustness and can be combined with any of the strategies yet to be described.

# 3.2.2 Model domain

If it is reasonable to assume that the ASR system is always used in the same noise environment, probably the simplest way of handling the problem is by training models using speech recorded in that particular environment (e.g., (Dautrich et al., 1983)), or speech corrupted by artificially adding the noise (e.g., (Gales, 1995)). These approaches have shown good results, but their use is limited to those situations where the speech to be recognized is always picked up in the same noise environment. Moreover, with this approach new models need to be trained for each new type of noise. Finally, artificially adding noise to clean recordings is only effective as long as the noise level under actual conditions is not so high that it gives rise to the Lombard reflex. This kind of spontaneous adaptation of the speech production enhances human intelligibility, but may very well harm ASR performance (Junqua, 1996).

If the noise is not easily predicted, but one can still obtain a reliable estimate of  $U(t, \omega)$ , one might want to use that estimate to adapt the observation distributions in the models trained on clean speech (Lee, 1998; Lee and Huo, 1999). Another example of a set of approaches developed for the model domain, is *predictive model combination*, *PMC*, also known as *parallel model combination* (Gales, 1998). In this case, the idea is to train separate models of noise and speech; if necessary, different types of noise can be modelled in parallel. During recognition the most likely combination of speech sounds and noise is computed. Searching the optimal path while using both noise and speech models leads to a three dimensional dynamic programming problem (Varga and Moore, 1990) with time, speech states and noise states as the three dimensions. If the noise can be described by an ergodic HMM, the three dimensional search problem can be converted into a conventional two dimensional search (Gales, 1998).

Despite the good results reported for different implementations of the PMC scheme (see for example (Gales, 1998)), such an approach is not always feasible. In particular, the usefulness of a PMC approach may be limited for two reasons, which are both growing more important with the increased use of mobile phones. Firstly, even if it is perfectly known beforehand what different noises can occur, the choice for the appropriate noise type will have to be made at recognition time. The decision will become more difficult as the number of different noise types known to the ASR system increases. Secondly, if the noise is time-variant, then the need will arise to continuously update the noise model on-line. Due to lack of observations, the noise model estimate may become poor to the extent that it limits the effectiveness of the compensation technique (Gales, 1998). These difficulties have inspired people to look for approaches that make less specific assumptions on how  $U(t, \omega)$  affects the features or the models. These are the approaches developed in the distance computation domain.

#### 3.2.3 Distance computation domain

The basic mechanism in an approach that does not rely on explicit estimates of the noise in terms of features or models consists of changing the similarity measurement between the trained models and the test utterance. The moment one realizes that feature values have an inherent uncertainty due to the presence of acoustic background noise, it is only natural to try to develop decision strategies that are primarily based on feature values that are least affected by noise characteristics. In (Lee and Huo, 1999) a number of such robust decision methods are discussed. These methods all attempt to account explicitly for the uncertainty in the feature values.

A somewhat different starting point is taken in the approaches based on *Missing Feature Theory* (MFT) (Cooke et al., 1996; Morris et al., 1998). According to Eq. 19 the signal spectrum at the ASR input can be considered as a mixture of a reliable component (i.e. the channel filtered original speech spectrum) and an unreliable component (i.e. the unknown noise contribution). Depending on the exact nature of the distortion some of the observed values in the acoustic feature vector may still be reliable, while other values may have become unreliable. In several recent proposals the key idea is to somehow disregard the unreliable information and base recognition on reliable information only. This idea can be pursued in different manners. First, let us suppose that one is working with acoustic feature vectors that are defined in the spectral domain, e.g., filter bank outputs. Then, if it can be assumed that  $U(t, \omega)$  takes non-negligible values only for a limited number of time frames t or a limited range of frequencies  $\omega$ , the marginalisation approach of MFT can be used (Cooke et al., 1996; Morris et al., 1998). With these assumptions it might be possible to explicitly detect all time-frequency regions where the observed feature values are dominated by  $U(t, \omega)$ , and either discard these features (Cooke et al., 1996; Dupont et al., 1997; Tibrewala and Hermansky, 1997; Lippmann and Carlson, 1997; Morris et al., 1998) or correct them in some way or another (Cooke et al., 1996; Morris et al., 1998; Dupont, 1998; Raj et al., 1998). Of course, the problem then immediately arises how corrupted values can be reliably detected. In the spectro-temporal domain this is not an easy task, although good progress was recently reported (Vizinho et al., 1999). If one is working with acoustic feature vectors defined in another domain (e.g., cepstra) then the detection task becomes even more intricate, because the components of the acoustic feature vectors that are significantly affected are not solely a function of the spectro-temporal characteristics of  $U(t, \omega)$ , but also of the transformations applied to the sequence of short-time spectra. We will elaborate this issue in more detail below.

Recently, a new way was suggested to handle contaminated feature values, which is not restricted to spectral features, and avoids the need to define a detector that is running independently from the decoder for identification of unreliable acoustic feature vector components. This idea, which was proposed in (de Veth et al., 1998c; de Veth et al., 2001), is yet another implementation of a method where the similarity measurement has been altered to cope with the noise. It focusses on the computation of the emission probabilities in the presence of disturbed acoustic feature vectors. This approach is based on the assumption that the statistical models built for clean speech are not proper models for observations obtained in the presence of acoustic feature vectors and each one of the candidate sequences of acoustic models (cf. Fig. 1), a situation is created in which unlikely feature values affect the search to a lesser degree. For convenience we use the term 'local distance function' when we refer to the mathematical expression used to evaluate the cost of assuming that a given sound segment pertains to a given HMM state.

If there is noise present at recognition time that was not present when the models were trained (i.e., in a mismatched training test condition), it is *a priori* evident that not all observations were actually seen in the training phase. Therefore, some part of the total probability mass is set apart to account for the unseen observations. For recognition, a new robust local distance function can then be determined by interpolating between the contributions of the cost for 'seen' and 'unseen' observations:

$$-\log[p(O)] = -\log[(1-\epsilon).p(O|seen) + \epsilon.p(O|unseen)],$$
(20)

where p(O) denotes the probability of the observation O, p(O|seen) is the probability of the observation according to the data seen during during training, p(O|unseen)the probability of the observation according the unknown process, and  $\epsilon$  the a priori probability that an observation is generated that was not seen during training. The idea of the robust local distance function in Eq. 20 is in fact an attempt to incorporate the well-known Tukey-Huber distortion model (Huber, 1981) in the recognition stage of an otherwise conventional HMM-based ASR (de Veth et al., 2001). What is essential here is the assumption that an observed event is the realization of a mixture of two processes. The first is the known process of which the parameters could be reliably estimated in the training phase, i.e., the process which produced the set of all seen observations. The second process is the one that produces all observations that were not seen in the training data. The only thing that is known about this second process is that some observations will be generated at recognition time that were not seen during training.

The idea that individual observations may originate from a mixture of a known and an unknown process plays an important role in the theories of Statistical Robustness (Huber, 1981) and Robust Statistical Pattern Recognition (Kharin, 1996). In the cases that we want to address (speaker independent recognition over the telephone), it is impossible to estimate the distortion distributions from the training speech. In addition, it is difficult to obtain a reliable estimate of the distribution of distortions from the unknown speech that is to be recognized. Under these two conditions we find ourselves in the situation (again) that it may be better to use an overly simplistic model, than to try to use a more sophisticated model. As we will see below, it is indeed possible to improve recognition performance based on an extremely simple assumption about the distribution of the observation values that were not seen during training.

In the remainder of this Section, we will first explain the robust local distance function in more detail. Next, we will introduce a topic that has not attracted much attention during the last decades, but that still might prove to be of considerable importance, viz. the way several transformations of the sequence of short-time spectra in the presence of additive noise (cf. Eq. 19) may affect the recognition result.

# **3.3** Disregarding unreliable information

#### **3.3.1** Robust local distance function

As stated before, the pattern match in state-of-the-art ASR systems is implemented as a search through frame-state space in the form of a dynamic programming algorithm (usually a Viterbi algorithm). For each acoustic feature vector, it is decided how each candidate optimal partial path so far is best extended with any of the HMM states that are candidates for extension. For each candidate optimal partial path, that state is selected which minimizes the path extension cost. For HMMs, this path extension cost is the combination of the emission cost of the candidate extension state and the transition cost for jumping to the candidate extension state (Rabiner, 1988). In what follows we will concentrate on the emission cost, since experience has shown that transition costs can actually be disregarded in a practical system without significant loss of recognition performance. Assuming that we really do not have any prior knowledge about the noise, which is not unreasonable when dealing with speech recognition over the telephone, one might reason as follows. An actually observed acoustic feature vector (or vector component) can be considered to be the realization of a mixture of two random processes: the known process as observed during training and the unknown process of all observations not previously seen. There is no need to explicitly determine by which of the two processes the observation was generated. It suffices to determine the emission cost due to the mixture of these two processes.

For an HMM state  $S_i$  that is described by a mixture of M Gaussian probability density functions the conventional local distance function  $d_{loc}$ , which we approximated to be equal to the emission cost, is described as

$$d_{loc}(\mathcal{S}_i, \mathbf{x}(t)) = -\log\{\sum_{m=1}^{M} w_{im} \prod_{k=1}^{K} G_{imk}(x_k(t))\},$$
(21)

where  $\mathbf{x}(t)$  denotes the acoustic observation vector at time t,  $w_{im}$  denotes the m-th mixture weight for state  $S_i$ , K denotes the dimension of the acoustic observation vector,  $x_k(t)$  the k-th component of  $\mathbf{x}(t)$ , and  $G_{imk}$  the k-th component of the m-th Gaussian probability density function for state  $S_i$ . The robust local distance function  $d_{robust}$  defined in (de Veth et al., 1998c; de Veth et al., 2001) is

$$d_{robust}(\mathcal{S}_i, \mathbf{x}(t)) = -\log\{\sum_{m=1}^{M} w_{im} \prod_{k=1}^{K} [(1-\epsilon)G_{imk}(x_k(t)) + \epsilon \hat{p}_0(x_k(t))]\}, \quad (22)$$

where  $\epsilon$  denotes the a priori probability that a feature value originates from the distribution of disturbed, unreliable speech values ( $0 \le \epsilon < 1$ ) and  $\hat{p}_0(x_k(t))$  denotes the unknown probability density function used to compute the probability for observing an outlier with value  $x_k(t)$ . It can be seen that Eq. 22 reduces to Eq. 21 if we choose  $\epsilon = 0$ .

Having reached this point, we still need to decide how the unknown process is best statistically described, where 'best' means optimal according to the principles of Robust Statistical Pattern Recognition. For the particular problem we study (i.e., how to make the computation of the local cost in the search robust) the best description of the unknown process is, as yet, an open question. However, this does not mean that one cannot make a sensible choice based on practical considerations. In (de Veth et al., 1998c; de Veth et al., 2001) it was proposed to model the unknown distribution as a uniform distribution, because this choice reflects our assumption best that we do not have any prior knowledge about the unknown process.

Another decision that remains to be made is how to choose the a priori probability  $\epsilon$  that a feature value originates from the distribution of values not seen during training. Without additional assumptions about the noise distortion there is no obvious way in which the 'optimal' value of  $\epsilon$  can be found. According to the experience gained so far, it appears to be reasonable to choose the Acoustic Backing-off parameter  $\epsilon$  such that the recognition performance in the matched training-test condition does not suffer too much, while in the mismatched condition the word error rate is maximally decreased (de Veth et al., 1999a; de Veth et al., 1999b).

We will now explain the effect of using the robust local distance function as defined in Eq. 22 and why the choice to model the unknown distribution as a uniform



value acoustic feature vector component

Figure 7. The contribution to the emission cost as a function of the observation value of one acoustic feature vector component for two competing states (indicated as 'i' and 'j'), when a conventional local distance function is used. Assuming that the current observation actually corresponds to state 'i', three observation values are considered: a reliable, undistorted observation value ('clean') and two different types of unreliable, distorted observation values ('d1' and 'd2'). For the conventional local distance function, the contribution to the emission cost due to a distorted value may lead to an unreliable assignment of the most probable state.

distribution is already convenient. In Fig. 7 the local distance functions corresponding to two competing, active HMM states (marked i and j) are shown for the conventional local distance computation. For illustration purposes, we have assumed that the emission probability density function  $p(x_k|S_i)$  is modeled as a single Gaussian. Then the local distance becomes a quadratic function of the difference between the value of the observed feature vector component and the mean value of the given distribution. We consider three different observation values: one undisturbed value corresponding to the clean condition (marked 'clean') and two different disturbed ones (marked 'd1' and 'd2'). We assume that the frame vector actually 'belongs' to state i. It can be seen that the contribution to the emission cost is lower for state i than for state j for the clean observation value. Now consider disturbed observation values 'd1' and 'd2'. In both cases the contribution to the emission cost for state *i* is (much) higher than the one for state *j*, thereby increasing the risk of recognition errors.

In Fig. 8 the same situation is depicted, but now the conventional local distance functions have been replaced by their robust versions. As can be seen, state i is being preferred over state *j* in the clean condition as before. However, for the distorted observation value 'd1' the contributions to the emission cost for states i and j have become



value acoustic feature vector component

Figure 8. The contribution to the emission cost as a function of the observation value of one acoustic feature vector component for the same two competing states 'i' and 'j' as shown in Fig. 7, now using a robust local distance function. With the robust local distance function, the contributions to the emission cost due to distorted feature values of type 'd1' become identical for the two competing states 'i' and 'j'. As a result, the assignment of the most probable state becomes independent of this type of distorted values. For a distortion of type 'd2', however, the robust local distance function is not effective.

identical. As a result the corrupted value will no longer favor the wrong state j. Obviously, it will not favor the right state *i* either. But if the corrupted value lies in the tail of all (or most) distributions for the active states, its contribution to the decision how to extend the candidate optimal partial path best is effectively canceled. If sufficient components of the acoustic feature vector of this frame contain uncorrupted values, they will discriminate between the active states and weigh in favor of the correct one. Obviously, this approach is not capable of removing the detrimental influence of distributional outliers of the type 'd2'. Here we are even more dependent on the presence of a sufficient number of undistorted values to compensate for the incorrect boost of the likelihood of state j.

The robust local distance function shown in Fig. 8 can also handle frames in which all values are corrupted, as long as the values are affected in the same manner as the 'd1' type of distortion. In this case, the emission cost for all competing states becomes essentially the same. When this happens, the frame makes no contribution to the decision of what is the best path and thus is effectively eliminated.

# 3.3.2 The effect of dispersion of unreliable information

As already indicated in the general scheme depicted in Fig. 1, in typical ASR systems the raw short-time spectra are not directly used for pattern matching. Most of the time, various normalization (e.g., gain normalization, channel normalization) and orthogonalization and dimensionality reducing transforms (e.g., Discrete Cosine Transform, Linear Discriminant Analysis) are applied. By using normalizing transforms, acoustic feature vectors are obtained that mainly represent the statistics of individual speech sounds and represent much less the variation due to differences in voice effort between different speakers or the variation due to different telephone channels. Orthogonalization transforms are used because they allow for more efficient modeling. For instance, only if the features are orthogonal, it is safe to assume that the covariance matrix is diagonal.

With clean speech data, normalization and orthogonalization transforms generally improve recognition performance significantly. However, a complication may arise when a subset of the components in the short-time spectrum are disturbed. In this case, corrupted values in a restricted number of short-time spectral components will be smeared out over the entire transformed vector. If this happens, the effectiveness of any strategy based on disregarding unreliable information might be jeopardized. This is readily illustrated for the case of MFT. The basic presupposition in MFT is that disturbances affect only part of the acoustic feature vector components and leave the rest intact. The idea of MFT is that recognition will be based only on those intact components. If some transform causes dispersion of the distortions over all acoustic feature vector components, none of the components are completely intact any more. The extent to which the effectiveness of MFT is undermined will then depend on how severely individual components are disturbed. In short, it is important to limit the spread of unreliable information in the acoustic feature vectors as much as possible, to keep the full effect of a strategy based on disregarding unreliable information.

#### 3.4 Connected digit recognition with additive band-limited noise

We studied the effect of the spread of unreliable information due to acoustic feature vector transformations and the effect of using a robust local distance function in the context of connected digit recognition over the telephone. In all experiments we started with mel-frequency log-energy coefficients as the basic representation of the short-time spectrum. These are the raw features. We compared the recognition performance for two types of acoustic feature representations. The first type of features are obtained by a full-smearing transformation of the raw features, i.e., a linear combinations of *all* raw features. For ease of reference, such feature representations are called F-type. The second type of feature representations are obtained by feature transforms of the raw features that only partly smear distortions over all feature vector components (P-type features). In particular, we used within-vector mean normalized mel-frequency log-energy coefficients (in short: F1) and mel-frequency cepstral coefficients (F2) and

compared these full-smearing transforms to sub-band mel-frequency cepstral coefficients (Okawa et al., 1998) (P1) and within-vector filtered mel-frequency log-energy coefficients (Nadeu et al., 1995b) (P2). To study the effect of the type of local distance function, we conducted two sets of experiments with connected digit recognition, one set with the conventional and the other with the robust local distance function. As a distortion we used additive band-limited Gaussian noise. The cut-off frequencies of the band-pass filter were chosen such that approximately one quarter of the energy bands that we used would be contaminated by noise  $(F_{low} = 395Hz \text{ and})$  $F_{high} = 880Hz$ ). We used three different signal-to-noise ratios of 20, 10 and 5 dBA respectively, i.e., both the speech and noise energy levels were weighted according to the A-scale (Hassall and Zaveri, 1979). The ten words used for the digits in Dutch were modeled using 3-state, context independent phone-based HMMs with 16 Gaussians per state. In all experiments reported here the robust local distance function was computed using  $\epsilon = 0.1$ . The uniform distribution that we used was defined independently for each component  $k, k = 1, \dots, K$  of the acoustic feature vector. Using all available observations in the training data, we determined a lower and upper bound  $(T_{k,low})$  and  $T_{k,high}$ ) such that 99.9% of all observations  $x_k(t)$  fell within the range between  $T_{k,low}$ and  $T_{k,high}$ . The uniform distribution for feature component k was defined to be equal to  $\frac{1}{T_{k,high}-T_{k,low}}$  inside this range and zero everywhere else. More details about the robust local distance function are given in (de Veth et al., 2001). Additional details about the experimental set-up can be found in (de Veth et al., 1999a; de Veth et al., 1999b). The recognition results using the conventional local distance function for the clean and noisy conditions are shown in Fig. 9. The results using the robust local distance function are shown in Fig. 10A, and the WER difference  $\Delta WER = WER_{robust} - WER_{conventional}$ is shown in Fig. 10B.

Looking at the clean conditions first, it can be seen that all four feature representations essentially perform at the same level and that recognition performance in the clean condition is affected only slightly when switching from the conventional to the robust local distance function. Focusing on the conditions where noise was added to the speech signals, two effects are clearly visible. Firstly, recognition performance is better for the two feature representations that only partially smear distortions (i.e., P1 and P2 (two rightmost bars)) than for the representations that smear distortions over all feature components (i.e., F1 and F2 (two leftmost bars)). This observation holds for the recognizer with the conventional as well as for the recognizer with the robust local distance function. Secondly, it can be seen that the recognizer based on the robust local distance function yields better results than the recognizer based on the conventional local distance function when noise is present in all cases, but one. The single exception occurs at SNR = 20 dBA for P1: The WER increases from 17.1% to 18.6% when switching from the conventional to the robust local distance function.

Given the data shown in Fig. 9 and Figs 10AB, two remarks are in place. Firstly, application of the robust local distance function in the clean condition consistently leads



**Figure 9.** Recognition results as a function of signal-to-noise ratio when using the conventional local distance function. F1: within-vector mean normalized Mel-frequency log-energy coefficients. F2: Mel-frequency cepstral coefficients. P1: sub-band Mel-frequency cepstral coefficients. P2: within-vector filtereded Mel-frequency log-energy coefficients.

to a slight loss of recognition performance. This is probably due to the fact that the distributions of the observations to be recognized in the clean condition are better represented by the distributions found during training than by the mixture of distributions used in the robust local distance computation. A mismatch between the mixture of distributions of observations could also explain the slight loss of recognition performance observed in one of the noise conditions. Secondly, the results show that performance improvements are observed for all types of features that were tested. In other words: Even for a feature representation that fully spreads spectrally local distortions over all feature vector components, the robust local distance function is capable of improving recognition performance. Apparently, the detrimental effect of the noise can be partially repaired by the robust local distance function, albeit that the improvement is not equally large for all feature types.

The results discussed in this Section cannot be readily generalized, because it must be expected that each specific noise type will affect different features differently. Consequently, it must be expected that the amount of success that our robust local dis-



Figure 10. A. Recognition results as a function of signal-to-noise ratio when using the robust local distance function. Same abbreviations as in Fig. 9. B. Corresponding  $\Delta$  WER results.

tance function can offer, will depend on a complex interaction between feature types, noise types and model characteristics. Additional research is needed to fully come to grips with this matter.

# 3.5 Discussion and conclusions

In this Section, we have presented a simple model of speech corrupted by additive noise that can be used as a framework to compare and understand several different approaches to making ASR more robust to noise. Noise robustness can be pursued in the acoustic feature domain, in the acoustic model domain, or in the distance measure domain. According to this scheme, we mentioned observation cleaning methods, of which spectral subtraction is the classic example. Next, we mentioned predictive model-based compensation (viz. by assuming that useful estimates of the parameters of the noise can be obtained, which in their turn can be used to adapt the models to better fit the conditions present in the new signal). As another approach in the model domain, we referred to work in model adaptation that attempts to make corrections based on observations as they are received at recognition time.

In our contribution, we have focused on conditions in which no dependable estimates of the noise can be made, so that we are left with the assumption that observed acoustic feature vectors can be considered as realizations of a mixture of two different processes. The first process is known and corresponds to the 'speech process' as observed during training. No knowledge about the second process is available. This unknown process corresponds to observations that were not seen during training. We have argued that this description allows to make a link with the theory of Robust Statistical Pattern Recognition (Kharin, 1996) and also to Missing Feature Theory (Cooke et al., 1996; Morris et al., 1998). These links deserve (and need) further research.

From a speech science point of view, two possible ways can be identified to extend the work on the robust local distance function. Firstly, our implementation of Missing Feature Theory might open new alleys towards the deployment of phonetic and auditory knowledge in automatic speech recognition. For example, it could be possible to introduce an estimate of fundamental frequency as one of the elements in the acoustic feature vector. Fundamental frequency is only defined for voiced speech segments and undefined for other signal portions. With Acoustic Backing-off, it should not be difficult to consider the value for fundamental frequency missing in signal portions that do not contain voiced speech. Secondly, the physical and perceptual basis under Missing Feature Theory might help to determine the best way for the incorporation of recent results from Robust Statistical Pattern Recognition in ASR.

We have also drawn attention to an issue which has not been widely discussed in the literature, viz. the potential interaction between transformations of the components of the short-time spectra and robustness against additive noise. With few exceptions additive noise will not affect all components of the short-time spectrum equally. We argued that a transformation may be suboptimal when it smears distortions which are local in the input short-time spectra over (almost) all components of the acoustic feature vectors. For Missing Feature Theory this is evident, since smearing violates the basic assumption underlying Missing Feature Theory, i.e., that part of the observation values are undistorted. The results from our experiments with feature representations that do not smear local distortion over the full feature vector have shown convincingly that it pays to minimize smearing. However, our research has also shown that it is not always straightforward to predict how a given distortion in the spectro-temporal domain will be smeared out in another domain under a given transformation. For instance, the fact that our robust LDF has a positive effect even for full-smearing features like within-vector mean normalized melfrequency log-energy coefficients and mel-frequency cepstral coefficients can at least in part be explained by the fact that not all transformed features have suffered equally from the smearing of the low frequency spectral distortion (de Veth et al., 1999b). More research is needed with respect to this subject.

As a final subject for further research, we recall that the robust local distance function in the form of Acoustic Backing-off is not capable to handle the 'd2' type distortions, shown in Figs 7 and 8. In general, the combination of the characteristics of the additive noise and the feature transform will result in a mixture of 'd1' and 'd2' type distortions. It is reasonable to expect that Acoustic Backing-off will be more effective if the proportion of 'd1' distortions in such a mixture is larger. However, it is still an open question how to handle a mixture with a large proportion of 'd2' type distortions to improve recognition robustness.

#### 4 Concluding remarks

In this chapter we have discussed two environmental factors which contribute to variation in speech signals and which therefore make automatic speech recognition difficult. The first factor is the effect of the transmission channel on the speech signal observed at the input of the ASR device; the second is the effect of additive noise. Both factors play their role in almost every recognition task, be it small vocabulary isolated word recognition or the recognition of unconstrained spontaneous speech.

Throughout this contribution, the discussion was based on a physical and mathematical model of the signals. It was argued that a model which encompasses all physical effects in great detail (including possible non-linear distortions) is far too complex to be handled. We have discussed how a simplified model can be adopted. Some simplifications are quite realistic, e.g., the assumption that the transmission channel is timeinvariant (or varies only very slowly) during a human-machine interaction session. Yet, in some specific situations the simplifications may become physically irrealistic. For instance, it is very unlikely that radio transmission errors in digital cellular networks are adequately represented by Eq. 19. However, it should be stressed that the simplifications we addressed in this contribution are motivated by the important finding that a simplistic model of which the parameters can be reliably estimated is always to be preferred over a physically more realistic model, if the parameters of the latter cannot be reliably obtained.

Another issue which has been emphasized throughout the paper is the interdependence of the modules of state-of-the-art ASR devices. Thus, an 'improvement' in one module, even one which is perfectly motivated by solid theory, may prove to deteriorate recognition accuracy, because it violates essential assumptions underlying other modules. This helps to explain why it has proven to be so difficult to harness conventional and recent knowledge from phonetics and auditory perception to improve ASR: Until we have a viable alternative for the Dynamic Programming search through a framestate space in which the frames constitute observations at equidistant time points with a single fixed frequency resolution, only the most basic phonetic and auditory perception knowledge can be brought to bear. We have illustrated this issue by means of RASTA filtering: It is precisely its relation to human temporal masking – and the attendant conflict with basic assumptions underlying HMM recognizers (like the assumption that speech can be modelled as a sequence of relatively invariant and static sub-word units) – which restricts the usefulness of classical RASTA to the realm of recognition based on wholeword or triphone units, and prevents its generalization to sub-word model systems based on units other than triphones.

Finally, we have pointed out how several different approaches to robust speech recognition can be unified or at least be related to one another. Again, the point of departure was a simplistic model, in which the signal at the input of an ASR device is considered as the sum of the 'clean' speech signal and some noise signal. Different approaches can be developed depending on the choice of the working domain (see Fig. 1): the acoustic feature domain, the model domain or the distance computation domain. Examples for the different domains are spectral subtraction, predictive model compensation, and Missing Feature Theory, respectively. We argued that the choice for any particular method depends on the assumptions about the parameters of the noise, and on the possibilities to reliably estimate these parameters. We have elaborated a recently emerged approach which has relations to the theory of Robust Statistical Pattern Recognition (Kharin, 1996) and more in particular to Missing Feature Theory (Cooke et al., 1996; Morris et al., 1998). Departing from the bottom line assumption that we have no prior knowledge about the noise, we have introduced 'Acoustic Backing-off' as a means for handling observations that are potentially corrupt and do not correspond to the distribution of observations seen during training. To that end we have introduced a new, robust local distance function. In doing so, we have uncovered a new issue, viz. the impact of feature transformations on the local distance function and the attendant search.

We hope that the presentations and discussions in this paper help to provide a framework to compare and unify the increasing stream of research papers on robust ASR. At the same time, it should help to guide future research and to focus it on those aspects which are most promising, given the full context of the models and assumptions implied in a speech recogniser. Finally, this framework should help to prevent disappointments by showing how local improvements can be turned counterproductive because of the way in which they violate critical assumptions in other components of a full-fledged ASR system.

# Acknowledgment

The contribution of Johan de Veth to this research was funded through the Priority Programme Language and Speech Technology (TST). The TST-Programme is sponsored by NWO (Dutch Organization for Scientific Research).

# References

- Aikawa, K., Singer, H., Kawahara H., Tohkura Y., 1993. A dynamic cepstrum incorporating time-frequency masking and its application to continuous speech recognition. In: *Proc. Internat. Conf. Acoust. Signal Speech Process.*, pp. 668–671.
- Anastasakos, A., Kubala, F., Makhoul, J., Schwartz, R., 1994. Adaptation to new microphones using tied-mixture normalization. In: *Proc. ARPA Spoken Language Techn. Workshop*, pp. 89–93.
- Atal, B., 1974. Automatic recognition of speakers from their voices. *Proc. IEEE*, 64, 460–475.
- Boda, P., de Veth, J., Boves, L., 1996. Channel normalisation by using RASTA filtering and the dynamic cepstrum for automatic speech recognition over the phone. In: *Proc. ESCA Workshop on the Auditory Basis of Speech Perception*, Keele, UK, pp. 317–320.
- Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.*, 27, 113–120.
- Cooke, M., Morris, A., Green, P., 1996. Recognising occluded speech. In: *Proc. ESCA Workshop on the Auditory Basis of Speech Perception*, Keele, UK, pp. 297–300.
- Dautrich, B., Rabiner, L., Martin, T., 1983. On the effect of varying filter bank parameters on isolated word recognition. *IEEE Trans. Acoust. Speech Signal Process.*, 31, 793–806.
- Drullman, R., Festen, J., Plomp, R., 1994. Effect of temporal envelope smearing on speech reception. J. Acoust. Soc. Amer., 95, 1053–1064.
- Dupont, S., Bourlard, H., Ris, C., 1997. Robust speech recognition based on multistream features. In: Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-a-Mousson, France, pp. 95–98.
- Dupont, S., 1998. Missing data reconstruction for robust automatic speech recognition in the framework of hybrid HMM/ANN systems. In: *Proc. Internat. Conf. Spoken Language Process.*, pp. 1439–1442.
- Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoust. Speech Signal Process.*, 29, 254–272.

- Gales, M., 1995. *Model-based techniques for noise robust speech recognition*. Ph.D. Thesis, Cambridge University.
- Gales, M., 1998. Predictive model-based compensation schemes for robust speech recognition. *Speech Communication*, 25, 49–75.
- Gish, H., Karnovsky, K., Krasner, M., Roucos, S., Schwartz, R., Wolf, J., 1985. Investigation of text-independent speaker identification over telephone channels. In: *Proc. Internat. Conf. Acoust. Signal Speech Process.*, pp. 379–382.
- Gish, H., Krasner, M., Russell, W., Wolf, J., 1986. Methods and experiments for textindependent speaker recognition over telephone channels. In: *Proc. Internat. Conf. Acoust. Signal Speech Process.*, pp. 865–868.
- Gong, Y., 1995. Speech recognition in noisy environments: a survey. *Speech Communication*, 16, 261–291.
- Haeb-Umbach, R., Beyerlein, P., Geller, D., 1995. Speech recognition algorithms for voice control interfaces. *Philips J. Res.*, 49, 381–397.
- Hassall, J., Zaveri, K., 1979. Acoustic noise measurements. Brüel & Kjær, Denmark.
- Hermansky, H., Morgan, N., Bayya, A., Kohn, P., 1991. Compensation for the effect of the communication channel in auditory-like analysis of speech. In: *Proc. Eurospeech*, pp. 1367–1370.
- Hermansky, H., Morgan, N., 1994. RASTA processing of speech. *IEEE Trans. Speech Audio Process.*, 2, 578–589.
- Hermansky, H., Pavel, M., 1995. Psychophysics of speech engineering systems. In: *Proc. Internat. Conf. Phon. Sc.*, pp. 3.42–3.49.
- Hermansky, H., 1996. Auditory modeling in automatic recognition of speech. In: *Proc. ESCA Workshop on the Auditory Basis of Speech Perception*, Keele, UK, pages in addendum.
- Hirsch, H., Meyer, P., Ruehl, H., 1991. Improved speech recognition using high-pass filtering of subband envelopes. In: *Proc. Eurospeech*, pp. 413–416.
- Huber, P., 1981. Robust Statistics. Wiley, New York.
- Hunt, M., 1978. Automatic correction of low-frequency phase distortion in analogue magnetic recordings. *Acoustic Letters*, 32, 6–10.
- Junqua, J.-C., Fohr, D., Mari, J.-F., Applebaum, T., Hanson, B., 1995. Time derivatives, cepstral normalisation and spectral parameter filtering for continuously spelled names over the telephone. In: *Proc. Eurospeech*, pp. 1385–1388.
- Junqua, J.-C., 1996. The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. *Speech Communication*, 20, 13–22.

- Junqua, J.-C., Haton, J.-P., 1996. *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. Kluwer Acad. Publ., Boston.
- Junqua, J.-C., 2001. Chapter to appear in: *Robustness in language and speech technol*ogy, G. van Noord, J.-C. Junqua (Eds.), Kluwer, Dordrecht.
- Kharin, Y., 1996. *Robustness in statistical pattern recognition*. Kluwer Acad. Publ., Dordrecht.
- Koehler, J., Morgan, N., Hermansky, H., Hirsch, H., Tong, G., 1994. Integrating RASTA-PLP into speech recognition. In: *Proc. Internat. Conf. Acoust. Signal Speech Process.*, pp. 421–424.
- Lee, C.-H., 1998. On stochastic feature and model compensation approaches to robust speech recognition. *Speech Communication*, 25, 29–47.
- Lee, C.-H., Huo, Q., 1999. Adaptive classification and decision strategies for robust speech recognition. In: *Proc. Workshop on Robust Methods for ASR in Adverse Conditions*, pp. 45–52.
- Lippmann, R., Carlson, B., 1997. Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise. In: *Proc. Eurospeech*, pp. 37–40.
- Little, J., Shure, L., 1993. *Matlab Signal Processing Toolbox Users Guide*. The Math-Works, Inc., Natick.
- Liu, F.-H., Moreno, P., Stern, R., Acero, A., 1994. Signal processing for robust speech recognition. In: *Proc. ARPA Spoken Language Techn. Workshop*, pp. 110–115.
- Lockwood, P., Boudy, J., 1992. Experiments with a non-linear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars. *Speech Communication*, 11, 215–228.
- Milner, B., Vaseghi, S., 1995. An analysis of cepstral-time matrices for noise and channel robust speech recognition. In: *Proc. Eurospeech*, pp. 519–522.
- Morris, A., Cooke, M., Green, P., 1998. Some solutions to the missing feature problem in data classification, with applications to noise robust ASR. In: *Proc. Internat. Conf. Acoust. Signal Speech Process.*, pp. 737–740.
- Nadeu, C., Paches-Leal, P., Juang, B.-H., 1995. Filtering the time sequence of spectral parameters for speaker-independent CDHMM word recognition. In: *Proc. Eurospeech*, pp. 923–926.
- Nadeu, C., Hernando, J., Gorricho, M., 1995. On the decorrelation of filter-bank energies in speech recognition. In: *Proc. Eurospeech*, pp. 1381–1384.
- Okawa, S., Bocchieri, E., Potamianos, A., 1998. Multi-band speech recognition in noisy environments. In: Proc. Internat. Conf. Acoust. Signal Speech Process., pp. 641-644.

- Orloff, J., Gillick, L., Roth, R., Scattone, F., Baker, J., 1994. Adaptation of acoustic models in large vocabulary speaker independent continuous speech recognition. In: *Proc. ARPA Spoken Language Techn. Workshop*, pp. 119–122.
- Rabiner, L., 1988. Mathematical foundations of hidden Markov models. In: *Recent advances in speech understanding and dialog systems*, NATO ASI Series, vol. F46, Springer-Verlag, Berlin, pp. 183–205.
- Rabiner, L., Schafer, R., 1978. *Digital processing of speech signals*. Prentice-Hall, Englewood Cliffs.
- Rahim, M., Juang, B.-H., 1996. Signal bias removal by maximum likelihood estimation for robust telephone speech recognition. *IEEE Trans. Speech Audio Process.*, 4, pp. 19–30.
- Raj, B., Singh, R., Stern, R., 1998. Inference of missing spectrographic features for robust automatic speech recognition. In: *Proc. Internat. Conf. Spoken Language Process.*, pp. 1491–1494.
- Schetzen, M., 1980. *The Volterra and Wiener theories of nonlinear systems*. Wiley, New York.
- Singer, H., Paliwal, K., Beppu, T., Sagisaka, Y., 1995. Effect of RASTA-type processing for speech recognition with speaking-rate mismatches. In: *Proc. Eurospeech*, pp. 487–490.
- Soong, F., Rosenberg, A., 1986. On the use of instantaneous and transitional spectral information in speaker recognition. In: *Proc. Internat. Conf. Acoust. Signal Speech Process.*, pp. 877–880.
- Steinbiss, V., Ney, H., Aubert, X., Besling, S., Dugast, C., Essen, U., Geller, D., Haeb-Umbach, R., Kneser, R., Meier, H.-G., Oerder, M., Tran, B.-H., 1995. The Philips Research system for continuous-speech recognition. *Philips J. Res.*, 49, 317–352.
- Tibrewala, S., Hermansky, H., 1997. Sub-band based recognition of noisy speech. In: *Proc. Internat. Conf. Acoust. Signal Speech Process.*, pp. 1255–1258.
- Varga, A., Moore, R., 1990. Hidden Markov model decomposition of speech and noise. In: Proc. Internat. Conf. Acoust. Signal Speech Process., pp. 845–848.
- de Veth, J., Bourlard, H., 1995. Comparison of hidden Markov model techniques for automatic speaker verification in real-world conditions. *Speech Communication*, 17, 81–90.
- de Veth, J., Boves, L., 1996. Comparison of channel normalisation techniques for automatic speech recognition over the phone. In: *Proc. Internat. Conf. Spoken Language Process.*, pp. 2332–2335.

- de Veth, J., Boves, L., 1997a. Channel normalisation using phase-corrected RASTA. In: *Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, pp. 119–122.
- de Veth, J., Boves, L., 1997b. Phase-corrected RASTA for automatic speech recognition over the phone. In: *Proc. Internat. Conf. Acoust. Signal Speech Process.*, pp. 1239–1242.
- de Veth, J., Boves, L., 1998a. Channel normalization techniques for automatic speech recognition over the telephone. *Speech Communication*, 25, 149–164.
- de Veth, J., Boves, L., 1998b. Effectiveness of phase-corrected RASTA for continuous speech recognition. In: *Proc. Internat. Conf. Spoken Language Process.*, pp. 963–966.
- de Veth, J., Cranen, B., Boves, L., 1998c. Acoustic backing-off in the local distance computation for robust automatic speech recognition. In: *Proc. Internat. Conf. Spoken Language Process.*, pp. 1427–1430.
- de Veth, J., de Wet, F., Cranen, B., Boves, L., 1999a. Missing feature theory in ASR: Make sure you miss the right type of features. In: *Proc. Workshop on Robust Methods for ASR in Adverse Conditions*, pp. 231–234.
- de Veth, J., Cranen, B., de Wet, F., Boves, L., 1999b. Acoustic pre-processing for optimal effectivity of missing feature theory. In: *Proc. Eurospeech*, pp. 65–68.
- de Veth, J., Cranen, B., Boves, L., 2001. Acoustic backing-off as an implementation of missing feature theory. Accepted for publication in: *Speech Communication*, 34 (3).
- Vizinho, A., Green, P., Cooke, M., Josifovski, L., 1999. Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: an integrated study. In: *Proc. Eurospeech*, pp. 2407–2410.
- Weintraub, M., Neumeyer, L., Digalakis., V., 1994. SRI November 1993 CSR spoke evaluation. In: *Proc. ARPA Spoken Language Techn. Workshop*, pp. 135–144.