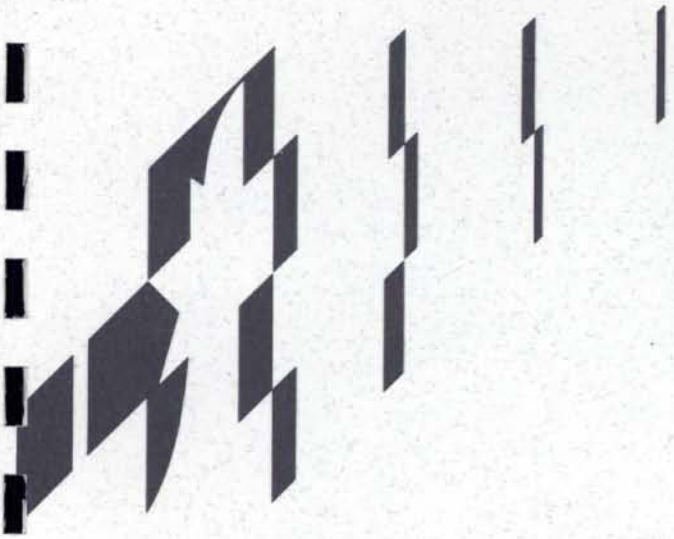

COMMUNICATIONS RESEARCH CENTER

CENTRE DE RECHERCHES SUR LES COMMUNICATIONS

SPEECH RECOGNITION IN NOISY ENVIRONMENTS: A SURVEY

Yifan Gong[1] and William C. Treurniet

Communications Canada

Canada

Broadcast Technologies Research Branch
Communications Research Centre
Department of Communications
Ottawa

SPEECH RECOGNITION IN NOISY ENVIRONMENTS: A SURVEY

Yifan Gong[1] and William C. Treurniet

CRC-TN 93-002

June, 1993

[1]Visiting scientist from CRIN/CNRS – INRIA-lorraine, France

Executive Summary

A good application of automatic speech recognition (ASR) would be the interface to information technology in mobile environments. Vehicle operators cannot safely operate keyboards and watch visual displays while driving the vehicle. ASR as a component of the user interface offers a less intrusive means for such interaction. However, ASR technology will not become practical in vehicles until recognition is robust in a changing auditory environment. For example, noise from the vehicle engine, the wind, and the tires all vary over time, and cause considerable difficulty for most current speech recognizers that are trained in a constant auditory environment.

This report summarizes a survey of the literature on speech recognition research that considers operating environments which may be different from the training environment.

The first author is a visiting scientist to CRC from CRIN/CNRS – INRIA-lorraine, France. His duties at CRC were to "initiate and perform research aimed at maintaining the performance of state of the art speech recognition technology in the unpredictable auditory environments found in automobiles". The work was performed under the general direction of W. C. Treurniet in the Broadcast Technologies Research Branch laboratory at CRC.

Abstract

The performance levels of most current speech recognizers degrade significantly when unpredictable noise occurs during use. Such performance degradation is caused by mismatches in training and operating environments. During recent years much effort has been directed to reducing this mismatch. This paper surveys research results in the area of single microphone noisy speech recognition classified in three categories: noise resistant features and similarity measurement, speech enhancement, and speech model compensation for noise. The survey indicates that the essential points in noisy speech recognition consist of incorporating time and frequency correlations, giving more importance to high SNR portions of speech in decision making, exploiting task-specific *a priori* knowledge both of speech and of noise, and including auditory models in speech processing. [1]

[1]Comments on several drafts of this paper by Mike Sablatash and Seymour Shlien of the Communications Research Center improved its quality and are very much appreciated.

Contents

List of Tables

1 Introduction

Speech recognition in controlled situations has reached very high levels of performance. For example, less than 0.5% word error was obtained in speaker-independent connected digit recognition using the TIDIGITS database [12], less than 1% error was obtained in speaker-dependent, isolated word recognition using a 20 thousand word vocabulary [19], and about 5% error was obtained in speaker-independent continuous speech recognition of the 1000 word vocabulary in the DARPA Resource Management task [5].

While most current speech recognizers give acceptable recognition accuracy for clean speech, performance degrades when they are applied to real situations, particularly in noisy environments. For example:

- The performance of a conventional word recognizer, trained with clean speech and giving 100% accuracy, can typically drop to 30% accuracy in a car traveling at 90 Km per hour [74].
- The 1% error rate of a system trained under quiet conditions increases to more than 50% in a cafeteria environment [19].

Environmental noise, therefore, has become one of the major obstacles to commercial use of speech recognition techniques.

Three phenomena are observed in noisy environments:

- Additive noise contaminates the speech signal and changes the data vectors representing frames of speech. For instance, white noise will tend to reduce the dynamic range, or variance, within the frame [7].
- Speaking in a noisy environment, where auditory feedback is obstructed by the noise, causes statistically significant articulation variability as the speaker attempts to increase the communication efficiency over the noisy medium. This phenomenon is known as the Lombard effect [41, 44, 61, 60].
- Spectral distortion occurs when the speech signal is convolved with an unknown linear system such as a different microphone or communication transmission line.

These phenomena produce serious mismatches between the training and recognition conditions that result in degradation in accuracy. Experiments have shown that a system trained under a given SNR (signal-to-noise ratio) usually gives poor recognition performance even when tested in a better SNR environment [62, 19, 82, 90, 38], due to the mismatch.

Consequently, all efforts in the field of noisy speech recognition have been directed to reducing the mismatch between training and operating conditions.

Some signal estimation techniques, such as basic Wiener and Kalman filtering, have succeeded in improving the SNR of noisy speech, but not necessarily the quality or the intelligibility of the speech [93, 84]. It seems plausible that, in optimizing the SNR, the spectrum of the speech signal is altered and some distortion is introduced which does not necessarily improve recognition accuracy.

A speech environment is characterized by a specific transmission line or noise condition, and a mismatching environment could be produced by either a microphone change or the introduction of a noisy background. Let s be the model of a recognition unit, e.g. a phoneme or word, e be an environment, and $q_e(s)$ be some quantity defined on s in the environment e. A transformation f is a mapping of quantities between two environments α and β, to be optimized to minimize operating environment error under some criterion:

$$q_\beta(s) = f(q_\alpha(s))$$

The problem of transformation is to find a function which decreases the mismatch between the training (reference) and operating (recognition) environments and thus improves the recognition accuracy in the operating environment.

The transformation can be either from the training environment to the operating environment (β = operating environment and α = training environment) or inversely. Based on the choice of the quantity q, two categories of transformations can be distinguished:

- Observation speech data transformation where, before recognition, speech data are transformed from a noisy environment into the environment in which models were trained.
- Speech model parameters transformation, where model parameters are transformed to match the noisy speech environment.

Therefore, the mismatch can be reduced in three basic ways:

- Assume that the system is noise independent, and use the same system configuration for both noisy and clean speech recognition. In this case, emphasis is on the search for noise resistant features.
- Transform noisy speech into a reference environment, and recognize it with a system trained in the reference environment. We call this scheme speech enhancement.
- Transform speech models created in the reference environment in order to accommodate the noisy environment and recognize noisy speech. This approach is called model compensation for noise.

Noisy speech recognition has become an important topic in many specialized journals and international conferences such as ICASSP, EUROSPEECH, and ICSLP. Although many publications on noisy speech recognition are available, no comprehensive description of recent research results has been found. This paper gives a survey of different single microphone, noisy speech recognition techniques classified in three categories by considering their initial objectives:

1. Identify noise resistant features and methods of similarity measurement,
2. Enhance noisy speech, and
3. Compensate speech models for noisy environments.

Because of the complexity of modern noisy speech processing strategies, techniques classified to the different categories may have some similarities. That is, different objectives sometimes lead to similar solutions.

Due to the wide coverage of theories involved in the field of noisy speech recognition, it is inappropriate to review in this paper the mathematical background of the methods described. For basic concepts, results, and references to other literature, the reader is referred to [78, 89, 21] for discussions of speech parameterization, to [98] for a discussion of HMM (hidden Markov modeling) of speech, to [72, 56] for discussions of ANNs (artificial neural networks) used as universal approximators, to [95] for a discussion of stochastic processes, and to [104] for a discussion of parameter estimation.

2 Noise Resistant Features and Similarity Measurement

2.1 Introduction

In a noisy environment, the distribution of parameters that give very good recognition results can be very sensitive to disturbances which introduce a mismatch between training and testing conditions. For example, it has been shown theoretically that the norm of cepstrum coefficients decreases as SNR decreases [77, 75]. Thus, for noisy speech, recognition accuracy drops drastically [76].

Speech recognition techniques described in this category focus on the effect of noise rather than on removal of the noise. They attempt to derive noise resistant feature parameters. Since feature parameters can only make sense if associated with some similarity measurement, similarity measurements to compare these parameters are

also studied. Often, the similarity measure is not independent of the representations being compared [7].

One of the advantage of these techniques is that no assumptions are made about the noise. On the other hand, this could be a shortcoming since it is impossible to make use of characteristics specific to a particular noise type.

2.2 Acoustic representation

In the DFT (discrete Fourier transform) domain, the combination of speech signal and most noise sources is additive and, therefore, relatively easy to process. However, it is currently believed that recognition performance is poorer with an observation probability distribution defined in the DFT domain than in the cepstrum domain [28]. Further, mel-scaled cepstrum coefficients are reportedly more resistant to noise than conventional LPC (linear prediction coefficients) cepstrum coefficients [73].

White noise corruption to speech signals reduces the norm of cepstral vectors. To compensate for this, a scale factor is incorporated into the cepstral Euclidean distance between a noisy test vector and a noise-free reference vector. From the orthogonality principle, the optimum value of the factor is the projection of the test vector onto the reference vector. Spectral analysis reveals [13] that the projection measure emphasizes the energy peaks of the spectrum and the difference between high energy frames involved in the distance.

Compared to Euclidean distance, the cepstral projection measure [77, 13, 14] has been shown to be noticeably less sensitive to noise. The lesser sensitivity is due to the fact that the angle between vectors is less affected by noise than their norms. Different cepstral lifters [49, 59], grand variance techniques which use a unique variance for all HMM states for a model [80], and smoothed variance measures [74] have also been found to reduce sensitivity to noise.

Also, a power or root operation can be applied to the spectrum of a noisy signal [68] instead of a logarithm operation. The resulting parameter is called a root compressed spectral density function [2] and was shown to be more resistant to noise than parameters resulting from the logarithmic transformation.

A SMC (short-term modified coherence) representation [76] obtained from an all-pole modeling of the autocorrelation sequence of a noisy signal followed by a spectrum shaping, can improve SNR by about 10-12 dB for noisy speech with a SNR range of 0-20 dB. When combined with a band-pass cepstral liftering technique, an improvement equivalent to 13 dB SNR was reported on an alpha-digits recognition task. An independent test confirmed that a given recognition rate was obtained by this method at about 15 dB SNR lower than the standard LPC-cepstrum method [81].

An HMM framework with norm equalization for observation densities was introduced

[57] in order to compensate for the shrinkage in the cepstral norm as noise is added. A performance improvement equivalent to 15-20 dB gain in SNR was obtained on a digit recognition test.

These feature representations and distortion measures are assumed independent of noise level and do not require specific training in the application environment.

2.3 Linear discriminant analysis

The IMELDA (Integrated MEl-scale Linear Discriminant Analysis) [55, 66, 54] was proposed for creating a noise resistant representation. IMELDA performs a linear transformation of a speech representation by minimizing within class differences and maximizing inter-class differences. The data are first transformed so that the average within-class covariance matrix is the identity matrix. Principal components analysis is then applied on the transformed between-class covariance matrix to select the direction of greatest variation. A subset of these principal components is used to form parameter vectors. Since the parameters maximize the average variation of phonemes between classes relative to the average within-class variation, IMELDA also improves clean speech recognition [99]. Since spectral contrast decreases as noise level increases, [7] further proposed normalization of cepstrum coefficients with respect to the frame standard deviation before processing with the IMELDA technique.

2.4 Spectrum model-based parameters

A model that is specific to speech sounds will accept signals with the properties of speech and reject perturbing noise, and thus be noise robust.

Assuming a certain spectrum structure of the speech signal, the LPC all-pole model has been used for improving the quality of the speech representation [67]. However, most LPC estimates degrade rapidly as the SNR decreases [92].

Under the constraint that the signal be an all-pole process, [45] proposed a maximum likelihood estimation of the speech waveform in additive noise. Although recognition accuracy was improved, the use of an LPC-based enhancement technique to bring the recognition performance to an acceptable level is questionable [40].

In constrained maximum *a posteriori* estimation of speech [43], some speech specific constraints such as the stability of an all-pole speech model, the relative position of poles, and the inertia of the vocal tract characteristics are exploited. The resulting speech has more reliable formant positions and reduced frame-to-frame pole jitter in noisy environments.

For LPC coefficients, a frequency-weighted Itakura spectral distortion measure was

proposed [103] which attempts to compensate for the bandwidth broadening effects due to the noise.

In [39], higher order derivatives of LPC with respect to NSR (noise-to-signal energy ratio) was used to compute an estimate of clean LPC. An estimate of the clean LPC vector was expanded in a Taylor series near the noisy LPC vector with respect to NSR, using up to fourth-order noisy LPC vector derivatives. The only information needed was the SNR, and no assumptions about the noise probability distribution were required. However, for an adequate approximation, the SNR must be large.

2.5 Auditory system inspired models

A number of authors have preprocessed speech data with computational models of the auditory system. These auditory models improve insensitivity to accompanying noise and, thus recognition accuracy [32]. However, severe performance degradation still remains for relatively intense noise environments (less than 10 dB SNR, for example).

Auditory models may incorporate lateral inhibition [100], which is a property of the auditory system. This phenomenon accounts for the observation that the perceived intensity of a test tone is a function of the frequency of an accompanying tone. When the two tones are similar in frequency, the perception of the test tone can be inhibited.

From the signal processing point of view, lateral inhibition results in the narrow-band SNR around spectral peaks to be higher than in spectral valleys. In [16] this effect is emulated by attenuating spectral valleys and emphasizing peaks, thus obtaining a significant improvement in SNR.

Auditory system models can be implemented as a wavelet transform followed by a compressive non-linearity [111]. The wavelet transform provides a multiresolution spectral representation of the signal, and spectral features were derived from locally averaged zero-crossing rates along the temporal axis. Lateral inhibition was obtained by computing the derivative with respect to the time and frequency axes of the transform. For speech degraded severely by noise, this representation preserved the spectrum structure of the speech signal significantly better then the power spectrum representation,

In [35] a computational auditory model based on the temporal characteristics of the information in the auditory nerve fiber firing patterns was presented. The model consists of a set of cochlear filters each followed by a zero-crossing detector and calculation of an interval histogram. The cochlear filters are equally spaced on a log-frequency scale. The output of the model is a frequency domain representation of the input signal in terms of the ensemble histogram of firing patterns. The output of the auditory model was converted to LPC coefficients which were input to a speech recognizer. Recognition experiments [36] showed that this representation was robust

with respect to noise contamination, and that the noise robustness was due to timing-synchrony analysis and not to the shape of the cochlear filters. Other experiments also report the ability of auditory models to enhance noisy speech signals [37].

In [33], a feedback model was proposed to simulate the efferent-induced effect on the cochlear and auditory-nerve fiber system. Compared to auditory systems without such feedback, the new system gave better resistance to degradation of recognition rate as noise level increased.

2.6 Discriminative similarity measures

A number of techniques exist to discriminate among different distributions of speech parameters vectors. The multilayer perceptron (MLP) classifier offers two important advantages over some other methods: no distribution is assumed for the probability density functions, and arbitrarily complex mappings can be created. In an experiment to classify single frames of noisy vowel data obtained from multiple speakers [94], an MLP classifier using cepstrum coefficients gave significantly better accuracy at all SNRs than a multivariate Gaussian maximum likelihood classifier and a k-nearest-neighbor classifier. Furthermore, the degradation in recognition accuracy with decreasing SNR occurred at a slower rate for the MLP than for the two other classifiers.

In evaluating the similarity between two sequences of parameter vectors, some vectors should be given more importance then others in order to achieve optimum global discriminability. For example, in E-set recognition, the dissimilarity at the /E/ part of the words is not discriminative and should not contribute to the recognition decision. In an experiment with a vocabulary-specific recognizer [3], two frames in an utterance were selected as discriminative parameter vectors, and were input to a MLP. The recognition rate was optimized for each sub-vocabulary by tuning the time position of one frame and the interval between the two frames. On a small vocabulary under clean, Lombard, and Lombard plus white noise conditions, the method gave substantially better recognition results than a continuous density HMM system.

2.7 Slow variation removal

Much additive noise as well as most channel distortions vary slowly compared to the variations in speech signals. Filters removing slow variations in the parameter feature vectors improve the recognition accuracy significantly [48]. The filtering may have different implementations and be in different parameter spaces, such as log-energy [51] or cepstrum [48].

RASTA (RelAtive SpecTrAl Processing) [51] consists of suppressing constant factors in each log spectral component of the short-term auditory-like spectrum. Each fre-

quency band is filtered by a filter with a sharp spectral zero at zero frequency. The average of each band is therefore zero. Since RASTA operates in the logarithmic domain, noise which is additive in the logarithmic domain such as slow-changing communication channel characteristics can be efficiently reduced. But noises that are additive in the time domain and therefore signal-dependent in the log domain cannot be removed efficiently [52]. RASTA has been shown to give good recognition accuracy with microphone variation [50].

Time derivatives (delta) of cepstra, also called dynamic features [29, 30] or first order regression features [4], improve recognition accuracy in noisy environments by removing slow variations. This technique is currently widely used in noisy speech recognizers as well as clean speech recognizers [47, 46, 48, 4].

In a speaker independent test with a 21 confusable words vocabulary, [48] has shown that suitable high-pass and band-pass filtering of log subband energies improves recognition of noisy Lombard speech, with little effect on recognition of clean speech.

Finally, it was observed in [28] that the common practice of preemphasizing the speech waveform degrades recognition accuracy in a noisy environment .

3 Speech Enhancement

3.1 Introduction

As a preprocessing step for recognition, speech enhancement techniques are intended to recover either the waveform or the parameter vectors of the clean speech embedded in noise [69]. These techniques make different uses of *a priori* information about the speech and the noise. The criteria used in speech enhancement techniques are based either on the probability of clean speech, the distortion between clean and recovered speech or directly related to speech recognition accuracy.

3.2 Spectral mapping

This approach to recovering clean speech from noisy speech observations directly exploits the one-to-one relationship between vectors in a clean speech environment and those in a noisy speech environment.

In [6, 84], linear multiple regression was proposed to map speech representations from a clean speech environment to a noisy environment. With this technique, transformation of reference tokens decreased the mismatch between training and operating conditions. It was shown that the transformation was superior to the spectrum subtraction technique [83].

More generally, arbitrarily complex transformations can be achieved using an artificial neural networks (ANN) such as the multi-layer feed-forward perceptron. ANNs have been used for noise reduction [107], particularly for vector sequence mapping [53], and mapping of a noisy magnitude spectrum to a clean magnitude spectrum [83]. Compared to a linear transformation, ANNs gave better results [83]. Due to their universal function approximation capability, very simple ANNs can be used to obtain reasonable results.

3.3 Noise subtraction

The spectrum subtraction approach to minimizing the effect of noise [18, 11, 74, 83] estimates the magnitude spectrum of clean speech by explicitly subtracting the noise magnitude spectrum from the noisy speech magnitude spectrum. The noise magnitude is estimated during nonspeech intervals of the voice communication process. In non-linear spectrum subtraction [74], the maximum of the local noise mean is used for subtraction.

The noise subtraction method assumes that the noise and speech are uncorrelated and additive. In that case, the magnitude spectrum of the noisy signal is the sum of the noise and the speech spectra. The method also assumes that the noise characteristics change slowly relative to those of speech signals, so that the noise spectrum estimated during a non-speech period can be used for suppressing the noise obtained during speech.

Subtraction techniques cannot be performed in the logarithmic spectrum domain where noise, even uncorrelated with the signal in the time domain, becomes signal-dependent. Also, negative spectral magnitudes can be obtained with subtraction, and usually *ad hoc* solutions to this problem are required.

If the parameter analysis is performed in spaces other than spectrum, the parameter can be transformed into the linear spectrum space where noise subtraction is performed. Spectral subtraction can also be performed on the inverse-transform of the LPC-cepstrum, as reported in [105].

3.4 Comb filtering

If the period of noisy voiced speech can be determined, then comb filters can be applied in the frequency domain to reduce the noise level. Comb filtering consists of multiplying the observation signal by a sequence of Dirac functions whose interval is the period of the signal. In the time domain, this operation is equivalent to averaging the signal waveform over several periods.

Comb filtering makes the assumption that noise is additive and short-time stationary.

Also, an accurate determination of the period of noisy speech is critical for the success of comb filtering. The filter is not applicable to speech segments with fast transitions, voiced fricatives, as well as unvoiced speech.

In [70], the effect of comb filtering on the SNR of speech in white noise was evaluated.

A speech signal can be represented as a sum of the harmonics of its fundamental frequency component. Under the assumption that the contaminating noise is Gaussian with known variance and that the amplitudes of the harmonics are independent, [64] has shown that the optimum estimation of speech may be considered to be a combination of comb filtering and Wiener filtering.

3.5 Bayesian estimation

Most of the techniques presented in this subsection were originally developed for speech quality improvement rather than for recognition. However, they may also serve as a pre-processing step for recognition systems.

If the clean speech is considered a function of random parameters of observed noisy speech, then the Bayesian approach can be applied to give an estimate of the parameters in order to obtain clean speech. Many cost functions can be used as criteria for the estimation. The most common ones are the squared error cost function which yields MMS (minimum mean square) estimation, and the uniform cost function which gives MAP (maximum a posteriori) estimation. The MMS estimate is the conditional mean of the parameter, given the observation. The MAP estimate is the value that gives the maximum conditional probability density function of the parameter, given the observation. In the case where observations and parameters are jointly Gaussian, the MMS and MAP estimates are identical. Minimum mean square error estimation of clean spectral channel energy from noisy channel energy has been applied to functions of DFT coefficients [25, 96].

Since the energies of individual channels are correlated and the correlations are speech-specific, a joint estimation of channel energies should be more robust than independent estimation. To partially incorporate the correlation, [27] conditioned the estimator on the total frame energy. Because more of the correlations between channels was incorporated, the new estimator significantly improved the quality of the estimate.

The estimator can be further enhanced by conditioning the channel energy on mixture models of the acoustic space [28]. This extension is based on the idea that the clean acoustic space can be divided into classes within which the correlation between different channels is significantly smaller than in the space as a whole, The resulting algorithm is called MMLSD (minimum mean log spectral distance). A theoretically weak point of this method is the assumption that the channels are statistically independent of each other. A Markov model was introduced to model time correlations

between adjacent frames by imposing continuity constraints on the estimate. Because of the complexity of speech compared to the simplicity of the model, the improvement was minor [28].

An HMM-based MMSE (minimum mean square error) estimator of clean speech was derived in [23]. The estimation introduces pairs of states of the models of the signal and of the noise. The estimator is a weighted sum of Wiener filters of noisy speech which are conditioned on the pairs of states. The weights are the probabilities of speech-noise composite HMM states given the noisy observations. Since HMMs are used in the selection of composite states, the estimate exploits information on the neighborhood of the analyzed vector.

In a more general signal mapping framework, [15] proposed a system for cleaning up speech which takes an (noisy) observation vector as input and a (clean) target vector as output. The relationship between input and output is, in general, a non-linear function. The input and output are each modeled as a set of random sources, with cross-correlations between them. A solution to the joint probability density function (pdf) of input observation, output target, and parameters of input and output random sources was proposed in [22] using the EM (expectation-maximization) algorithm. Based on the pdf, the desired output vector is computed as the conditional expectation of the output, which is a minimum mean square estimate of clean speech. In [17], noisy speech as input sources was converted with this method to auditory spectra [16] to obtain resistance to noise and less dependence on noise level relative to conventional spectral mapping.

Using the MAP approach, a sequence of speech-noise states via HMM was estimated, and the states were used to construct Wiener filters [26]. Clean speech was modeled as a hidden Markov process with mixtures of Gaussian AR (autoregressive) output processes. Noise was assumed to be additive and independent of the signal, and was modeled as a Gaussian AR vector. The enhancement of noisy speech was performed iteratively by alternating

- Estimation by a Viterbi decoder of the most probable sequence of states and mixture components.
- Estimation of clean speech by application of state-dependent, thus time-varying, Wiener filters.

MAP estimation theory was also applied in [23].

Under a dynamical system framework, the alternate Viterbi decoding – Wiener filtering of [26] is generalized to the alternate Viterbi decoding – Kalman filtering speech enhancement scheme [24].

3.6 Template-based estimation

Within-frame speech-specific frequency dependencies of speech can be explicitly exploited by the frame template-based approach. The processing involves finding the best sequence of clean speech templates given the noisy speech data. The basic property of the template-based approach is the restriction of the signals to be estimated, to a parameter subspace defined by templates and the combination coefficients. The advantage of using templates is that the output is almost noise free insofar as the noise influence is converted into incorrect estimation of templates.

Templates can be either phoneme-based [38], VQ (vector quantization)-based [58, 92], or a linear combination of sine-waves [97] for temporal signal estimation. The mostly linear combination of templates can be derived from the Gaussian assumption [97], from fuzzy vector quantization [38], from averaging several templates closest to the noisy vectors [58] or just taking the closest vector under some distance metric [92].

In [38], clean speech estimates were given by linear combinations of clean speech templates. The templates were extracted from a set of phonemes. The combination coefficients were based on the similarity between the noisy vector and noisy speech templates. The templates were obtained from labeled phonemes in pre-specified utterances. On a 200-word vocabulary, under 10 dB SNR, a 90% recognition rate was obtained using a recognizer that gave 95% for clean speech. The recognition system used phonemes as basic recognition units.

In [58], to achieve speech signal restoration, a given clean speech reference template was computed as the average of several clean speech templates. These templates were chosen from a set of templates so that their noisy correspondence gave the smallest distortion to the noisy observation. The set of templates was obtained by applying the Lloyd algorithm [71] to clean speech, and using a likelihood ratio or a truncated cepstral distance distortion measure. It was shown experimentally that, under 14 dB SNR, an improvement of 10 dB SNR was achieved.

For the purpose of improving noisy speech intelligibility, it was proposed in [92] to resynthesize speech from VQ templates. Formant distance was used as the similarity measure. The output of such processing is noise-free speech, with the remaining degradation appearing as a spectrum mismatch.

A least square error estimate of the desired speech waveform in the presence of noise was proposed in [97]. A speech waveform was represented as a linear combination of sine wave templates. The combination coefficients were roughly inversely proportional to the distance of each scaled template to the observed speech data. The weighting scheme was similar to that used in fuzzy vector quantization [108].

Usually template-based estimation methods are not formulated on a probabilistic basis. However, under some assumptions about the signal and noise, the results are similar to those of some methods in subsection 3.5, since templates are equivalent to

parameters describing observation probabilities.

4 Model Compensation for Noise

4.1 Introduction

Rather than attempting to derive a best estimate of the speech signal, approaches using model compensation for noise allow for the presence of the noise in the recognition process itself.

HMMs provide a formal mathematical framework for modeling temporal and spectral variability in speech signals. In this framework, solutions to parameter estimation and speech recognition are formulated as widely used and computationally attractive algorithms. Model-based recognition such as HMM offers the possibility to exploit a specific model of clean speech obtained from a training process, and then changing the model parameters to accommodate noisy speech. That is, noisy speech data is used to adapt speech model parameters such as the mean and variance of a Gaussian distribution, in order to compensate for the discrepancy between training and recognition conditions.

Such adaptive schemes are potentially able to deal with noisy environments which are not presented in the training phase and/or which are time varying.

4.2 Decomposition of hidden Markov models

This is a generalization of hidden Markov modeling to optimal decomposition of simultaneous processes [110]. With this technique, noisy speech signals can be modeled by an $N \times M$ state HMM, where N is the number of states for clean speech and M is for the noise process. The standard Viterbi algorithm searches for a best path among N (states) $\times T$ (observations), which can be extended to $N \times M \times T$, where M is the number of states of a noise HMM. The transition probabilities of the signal model and of the noise model are trained separately. The output probabilities of the two models are combined under some assumptions to give the probability of observing the input sequence. The combination is dependent on the feature parameter space of the signals. The decomposition allows simultaneous recognition of both signal and noise. A similar approach has been used in [23] where the purpose of introducing pairs of states was speech enhancement.

Using a filter bank [110], under a simplified model of output probabilities and simple models of noise, the signal decomposition model provided significant recognition performance improvement for both stationary and highly non-stationary noise. In

stationary pink noise, good recognition accuracy was obtained down to a SNR of -3 dB.

In the same framework, a mel-cepstrum model combination was studied in [31], where model parameters in the cepstral domain are transformed into the linear spectral energy domain, adapted, and then transformed back to the cepstral domain. The recognition rate was shown to be better than that of [110]. In an evaluation using spoken digits, this technique outperformed a noise masking technique, especially for SNRs down to -3 dB [80].

Model decomposition provides a framework for accommodating independent concurrent processes. Potentially this approach can deal with non-stationary interfering signals, since the noise models can accommodate very complex noises with fast changing and impulsive statistical characteristics.

4.3 State-dependent filtering in hidden Markov models

Direct application of Wiener filtering to noise cancellation for noisy speech is limited by the non-stationarity of speech signals. HMMs automatically divide speech into quasi-stationary segments corresponding to the model state. This property of HMMs is exploited in [8] to implement noise cancelling filters within a speech recognition process. Under this scheme, in each state for each filter channel, an optimum FIR Wiener filter is attached as an additional parameter of the model. During recognition, a filtered estimate of clean speech is calculated on the basis of a sequence of noisy input vectors. The estimate is used to compute the output probability of the state.

In the frequency domain, Wiener filtering corresponds to a multiplication of the signal spectrum and the filter frequency response. In the cepstral domain, this is equivalent to an additive operation due to the logarithmic scaling involved. In [9], a HMM state-dependent Wiener filter was implemented to additively correct cepstral observation vectors prior to the calculation of the output probability for that state. Cepstral correction by Wiener filtering was also used in a template-based dynamic programming recognition system [10], where the filters were frame-dependent.

State-dependent Wiener filtering is similar to some of the methods described in subsection 3.5, where the purpose of Wiener filtering was speech enhancement rather than noise cancellation during recognition.

4.4 Adaptation of duration models

The Lombard effect produces changes of speech in both spectral characteristics and timing structures of acoustic events. Both types of changes contribute to the mismatch between training and operating conditions and, thus, to the degradation of a

recognizer's performance. Most noisy speech recognizers deal only with the first type of mismatch but not with the second, which limits the improvement in performance.

In [102], it was proposed to re-estimate the parameters of duration models of phonemes trained in a clean environment, using a few observations of Lombard speech. The parameters of the models were considered as random variables estimated under the MAP criterion. It was shown that the resulting mean of a phoneme duration model approaches the one from the clean environment if there are few adaptation data, and approaches the MLE (maximum likelihood estimation) of the mean of the adaptation data when the quantity of adaptation data is large. Experiments showed that when the mismatch in spectral characteristics was very great, duration adaptation was very efficient in improving recognition performance.

4.5 Adaptation of hidden Markov models

When HMM-based speech recognizers are trained with clean speech data, performance is degraded when test data is contaminated with noise. Some attempts have been made to use a small amount of the noisy speech to adapt the HMM model parameters to the noisy environment.

For discrete HMM speech recognizers, instead of using a simple Euclidean distance measure, [86] introduced mixtures of label prototypes. The pdf of a prototype was based on a weighted sum of the pdfs of the noise vector and the clean speech vector. The energy of each frequency band was assumed to be the maximum of clean speech and noise energies. The parameters of the noise was estimated during recognition by fitting a Gaussian distribution to the noise.

In [83], a linear transformation diagonal matrix was proposed to directly adapt – rather than transforming the observation speech signal – the state output probability parameters of a HMM recognizer. The parameters included means and variances of Gaussian distributions.

In [1], an iterative codeword-dependent additive normalization of cepstral parameters was proposed, which improved recognition rates for noisy speech.

The Bayesian learning procedure has been used for adapting Gaussian state observation densities of a continuous density HMM [65]. Using two or three repetitions of each word, the method improved a severe mismatch between the training and testing recording conditions. Because the Bayesian procedure adapts a trained parameter set to a new environment, training data is used more efficiently than with conventional codebook mapping approaches.

Since the nature of the adaptation problem is to find a mapping between two spaces, techniques originally proposed for speaker-adaptation such as vector quantization codebook mapping [101, 87], fuzzy vector quantization-based spectral mapping which

has been shown to give better spectrum distortion than neural network based nonlinear mapping [88], vector field smoothing [91], observation distribution adaptation [106], mixture coefficients modification [79] and Bayesian learning for Gaussian mixture HMM state observation densities [34], can be applied to noisy speech recognition.

4.6 Minimum error training

Conventional HMMs are trained using the maximum likelihood estimation (MLE) scheme, which maximizes the probability of the training data given a model. Since a model of a given class is trained independently of others, discrimination between different classes cannot be optimized during the training. Minimum error training, also called corrective training or discriminative training, was proposed to overcome this deficiency by maximizing the difference between the probability of the correct class and the probability of the most probable incorrect reference class. This leads to the minimization of misclassification. Generally the parameters of the models are initialized by MLE training before minimum error training begins. This technique has been shown to improve the recognition rate for clean speech.

In [82], minimum error training with noisy speech was used to train HMMs initialized by MLE using clean speech. The minimum error training was performed on the mean vector of the Gaussian pdf of the most probable mixture component of each state. Significant improvement in recognition accuracy was reported.

In [90], both the mixture coefficients and the mean vectors of a continuous density HMM system were adjusted by minimum error training. The results confirmed that the technique helps to maintain recognizer accuaracy as noise is added to the environment.

4.7 Noise masking

Noise masking [63] is the psychological phenomenon of reduction of perceptibility of a signal in the presence of noise. The effect of this phenomenon can be emulated to improve noisy speech recognition performance. Masking was simulated in [109] by a masking algorithm for spectral energy. For both speech model parameters and input speech, the value of each frequency band was replaced by the noise mean if the latter was greater than the former. For speaker-dependent digit recognition in pink noise, robust accuracy was reported for SNR down to 3 dB.

A cepstral equivalent of this algorithm was also described [80], which maintains a log energy version of the models for the masking operation and transforms the masked models to the cepstral domain.

4.8 Training data contamination

As a special case of model compensation, another strategy is to add noise to the training tokens [20, 85]. With this technique, the mismatch between training and operating environments will totally disappear. Use of noise contaminated data to train a system can dramatically improve recognition accuracy under that specific training condition. The reported results are better than those of other more sophisticated processing techniques such as Kalman filtering and spectral transformation [84].

Using a HMM-based recognizer, [82] compared MLE training using clean speech followed by minimum error training using speech with 10 dB of noise on the one hand, and MLE training using both the clean speech and the noisy speech on the other hand. When tested with different SNRs and noise types, the recognition rates for the two training schemes were similar, with the first scheme giving slightly better performance.

Evidently, however, the technique cannot cope with the Lombard effect. Another problem for this technique is that the noise to add to the training data is not always available. Moreover, since no noise model is incorporated and since the recognition accuracy is only optimized to the intensity characteristics of the training noise, recognition performance could be sensitive to noise level [62].

Since no mismatch is involved with the training data contamination method, many authors refer to recognition accuracy obtained using this technique as a performance reference to which to compare noisy speech recognizers.

5 Conclusion

Speech recognition in noise involves a large variety of knowledge at every level of processing. Due to the complex nature of noisy speech recognition, data for accurate comparative performance evaluation of the techniques are unavailable. Table 1 gives an incomplete overview of achieved performance improvement in recognition of speech signals corrupted by additive Gaussian noise. In the table, the recognition results are given, in percent correct, for system performance in 10 dB SNR. The column labeled *comparison* gives the recognition results for the same data when the same system was used without the proposed noise resistant feature. For each method, the recognition rate for clean speech is also given. The number in round brackets gives the perplexity of the grammar used, if applicable. For equal SNR, it was observed that for types of noise other than Gaussian, less degradation generally resulted [42, 10, 82].

Since the purpose of noisy speech recognition techniques is to reduce the mismatch between training and testing conditions, many authors aim at achieving the recognition accuracy comparable to that obtained by training the recognizer with noisy

method	vocabulary	speaker	clean	10dB	comparison
HMM lpc-mel [57]	10 digits	multi	100.	30.5	–
HMM vector equalization [57]	10 digits	multi	98.3	77.3	30.5
modified EIH [36]	39 words	single	95	70	25
SMC [76]	alphabet	depend	93.2	73.2	35.4
SMC [76]	10 digits	depend	99.2	98.3	39.8
IMELDA [7]	alphabet	independ	89.8	39.0	12.1
LC-MBCE-HMM [42]	35 words	multi	96	62	8
Constrained MAP [43]	20 words	single	88	34.5	5
Projection [13]	34 words	single	95	88	40
Projection [14]	10 digits	independ	98	83.6	38.4
ME training [82]	100 words	independ	99.1	94.9	66.3
MMLSD-VQ [28]	1K words (60)	independ	92	78.9	8
base-transform [38]	206 words (206)	single	95.0	89.8	20.4

Table 1: Recognition accuracy in 10 dB Gaussian additive noise of some speech recognition systems.

speech. Presently, the gap can be very small, about 1-2 percent.

In general, if the statistical properties of the noise is known, it is easier to obtain an improvement in recognition accuracy using a transformation-based technique than a feature-similarity-based technique. However, feature-similarity-based techniques, described in Section 2, are usually more robust over varying SNR, with decreasing recognition accuracy as noise level increases. Unfortunately, there is no theoretical justification to claim that the Lombard effect, which involves changes in the articulatory system and speech rhythm, can be directly handled with the methods described in Section 2. Techniques in Sections 3 and 4 usually give optimum recognition results for the noise variety and level to which the system was trained. When operating conditions change far enough from the training condition, the recognition rate will decrease.

In speech recognition applications, methods described in Section 2 are computationally efficient. Those in Section 3 and Section 4 usually require collecting noisy data and training the system. Combinations of different classes of techniques may further improve a recognition system's performance.

Many methods assume the statistical independence of the speech signal and the noise. Although this assumption simplifies treatments and results in computationally attractive algorithms, it inherently limits the performance for real applications. For instance, the Lombard effect, if it is considered as noise, is not signal independent.

The key problem in noisy speech recognition is to define a mathematically attractive optimal criterion which takes into account the following:

- Incorporating inter-frame correlations including correlations across vector sequences over time. The hidden Markov modeling provides a good framework.
- Incorporating frequency correlations including the correlation across components of representation vectors via LPC, AR, or other techniques such as template-based approaches.
- Giving more importance to high SNR portions of speech in decision making. The portion could be either a time interval or a narrow band frequency range.
- Modeling explicitly *a priori* information about speech and noise by exploiting models of both speech and noise. State decomposition strategy is a promising approach.
- Including auditory models in speech processing.

Two major issues remain. The first is insensitivity to different noise levels, which requires the generalization of training results for one SNR to other SNR conditions. Current transformations depend strongly on the noise levels for which they are trained. In most cases, when a speech recognizer performs well with noisy speech, performance drops when recognizing clean speech. The second problem is dealing with non-stationary noise environments. Algorithms which automatically optimize recognition accuracy while the recognizer is in use need to be developed further.

The most appropriate algorithms for a given situation are dependent on the research objectives as well as assumptions about the statistical properties of relevant variables. Also, the outcomes of the various methods for dealing with noise may be different depending on the characteristics of the noise. It is therefore important to have a suitable database for validation of a given method, since the type of noise and criterion for optimization could influence the result. The validation criteria for noisy speech recognition systems can be very complex. However, they should provide information on the improvement in recognition accuracy as compared to a standard noise level (e.g., 10 dB SNR), and on the recognition accuracy as a function of SNR relative to the SNR under which the system was optimized.

References

[1] A. Acero and R. Stern. Environmental robustness in automatic speech recognition. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 849–852, April 1990.

[2] P. Alexandre, J. Boudy, and P. Lockwood. Root homomorphic deconvolution schemes for speech processing in car noise environment. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 99–102, 1993.

[3] Y. Anglade, D. Fohr, and J.-C. Junqua. Speech discrimination in adverse conditions using acoustic knowledge and selectively trained neural networks. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 279–282, 1993.

[4] T. H. Applebaum and B. A. Hanson. Regression features for recognition of speech in quiet and in noise. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 985–988, 1991.

[5] X. Aubert, R. Haeb-Umbach, and H. Ney. Continuous mixture densities and linear discriminant analysis for improved context-dependent acoustic models. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 648–651, 1993.

[6] L. Barbier and G. Chollet. Robust speech parameter extraction for word recognition in noie using neural networks. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 145–148, 1991.

[7] D. C. Bateman, D. K. Bye, and M. J. Hunt. Spectral contrast normalization and other techniques for speech recognition in noise. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing 1992*, volume I, pages 241–244, San Francisco, U.S.A., March 1992.

[8] V. L. Beattie and S. J. Young. Noisy speech recognition using hidden Markov model state-based filtering. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 917–920, 1991.

[9] V. L. Beattie and S. J. Young. Hidden Markov model state-based cepstral noise compensation. In *International Conference on Speech and Language Processing*, volume I, pages 519–522, Banff, Alberta, Canada, October 1992.

[10] A. D. Berstein and I. D. Shallom. An hypothesized Wiener filtering approach to noisy speech recognition. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 913–916, 1991.

[11] S. F. Boll. Suppression of acoustic noise in speech using spectral substraction. *IEEE Trans. on Acoust., Speech and Signal Processing*, ASSP-27(2):113–120, April 1979.

[12] R. Cardin, Y. Normandin, and E. Millie. Inter-word coarticulation modeling and MMIE training for improved connected digit recognition. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 243–246, 1993.

[13] B. A. Carlson and M. A. Clements. Application of a weighted projection measurement for robust hidden Markov model based speech recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 921–924, Toronto, Canada, May 1991.

[14] B. A. Carlson and M. A. Clements. Speech recognition in noise using a projection-based likelihood measure for mixture density HMM's. In *Proc. IEEE Int. Conf. on*

Acoustics, Speech and Signal Processing 1992, pages 237–240, San Francisco, U.S.A., March 1992.

[15] Y. M. Cheng, D. D'Shaughnessy, and P. Mermelstein. Statistical signal mapping: A general tool for speech signal processing. In *Proc. of* 6^{th} *IEEE Workshop on Statistical Signal and Array Processing*, pages 436–439, 1992.

[16] Y. M. Cheng and D. O'Shaughnessy. Speech enhancement based conceptually on auditory evidence. *IEEE trans. on ASSP*, 39(9):1943–1954, September 1991.

[17] Y. M. Cheng, D. O'Shaughnessy, and P. Kabal. Speech enhancement using a statistically derived filter mapping. In *Proc. of Int. Conf. on Spoken Language Processing 1992*, volume I, pages 515–518, Banff, Alberta, Canada, October 1992.

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

[19] S. Das, R. Bakis, A. Nadas, D. Nahamoo, and M. Picheny. Influence of background noise and microphone on the performance of the IBM TANGORA speech recognition system. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 71–74, 1993.

[20] B. A. Dautrich, L. R. Rabiner, and T. B. Martin. On the effect of varying filter bank parameters on isolated word recognition. *IEEE Trans. on Acoust., Speech and Signal Processing*, ASSP-31:793–806, August 1983.

[21] S. B. Davis and P. Mermelstein. Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech and Signal Processing*, ASSP-28(4):357–366, August 1980.

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[23] Y. Ephraim. A Bayesian estimation approach for speech enhancement using hidden Markov models. *IEEE Trans. on Acoust., Speech and Signal Processing*, ASSP-40(4):725–735, April 1992.

[24] Y. Ephraim. Speech enhancement using state dependent dynamical system model. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing 1992*, volume I, pages 289–292, San Francisco, U.S.A., March 1992.

[25] Y. Ephraim and D. Malah. Speech enhencement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. on Acoust., Speech and Signal Processing*, ASSP-32:1109–1112, Dec 1984.

[26] Y. Ephraim, D. Malah, and B.-H. Juang. On the application of hidden Markov models for enhancing noisy speech. *IEEE Trans. on Acoust., Speech and Signal Processing*, ASSP-37(12):1856–1855, Dec 1989.

[27] A. Erell and M. Weintraub. Energy conditioned spectral estimation for recognition of noisy speech. *IEEE Trans. on Speech and Audio Processing*, SAP-1(1), Jan 1993.

[28] A. Erell and M. Weintraub. Filterbank-energy estimation using mixture and Markov models recognition of noisy speech. *IEEE Trans. on Speech and Audio Processing*, SAP-1(1), Jan 1993.

[29] S. Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE trans. on ASSP*, 34(1):52–59, February 1986.

[30] S. Furui. On the use of hierarchical spectral dynamics in speech recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing 1990*, pages 789–792, Albuquerque, New Mexico, USA, April 1990.

[31] M. J. F. Gales and S. Young. An improved approach to the hidden Markov model decomposition of speech and noise. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 233–236, U.S.A., April 1992.

[32] Y. Gao, J.-P. Haton, and Y. Gong. Noisy speech recognition tested on continuous speech recognition system. In *Proc. of ESCA Workshop on Speech Processing in Adverse Conditions*, France, Nov 1992.

[33] Y. Gao, T. Huang, S. Chen, and J.-P. Haton. Auditory model based speech processing. In *International Conference on Speech and Language Processing*, volume I, pages 73–76, Banff, Alberta, Canada, October 1992.

[34] J. L. Gauvain and C. H. Lee. Bayesian learning for hidden Markov model with Gaussian mixture state observation densities. *Speech Communication*, 11:205–213, 1992.

[35] O. Ghitza. Auditory nerve representation as a front-end for speech recognition in a noisy environment. *Computer, Speech and Language*, 2, 1987.

[36] O. Ghitza. Robustness against noise: the role of timing-synchrony measurement. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, pages 2372–2375, 1987.

[37] O. Ghitza. Auditory neural feedback as a basis for speech processing. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 91–94, 1988.

[38] Y. Gong. Base transformation for environment adaptation in continuous speech recognition. In *Proceedings of European Conference on Speech Communication and Technology*, Berlin, Germany, September 1993.

[39] C. Guan, Y. Chen, and B. Wu. Direct modification on LPC coefficients with application to speech enhancement and improving the performance of speech recognition in noise. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 107–110, 1993.

[40] J. H. Hansen and M. A. Clements. Iterative speech enhancement with application to automatic speech recognition. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 561–564, 1988.

[41] J. H. L. Hansen. *Analysis and Compensation of Stressed and noisy speech with application to robust Automatic recognition.* PhD thesis, Georgia Institute of Technology, 1988.

[42] J. H. L. Hansen and O. N. Bria. Improved automatic recognition of speech in noise and Lombard effect. In J. Vandewalle, R. Boite, M. Moonen, and A. Oosterlinck, editors, *Signal Processing VI: Theories and Applications*, pages 403–406. Elsevier Science Publishers B. V., 1992.

[43] J. H. L. Hansen and M. A. Clements. Constrained iterative speech enhancement with application to speech recognition. *IEEE Trans. on Signal Processing*, 39(4):795–805, April 1991.

[44] J. H. L. Hansen and O. N. Oria. Lombard effect compensation for robust automatic speech recognition in noise. In *International Conference on Speech and Language Processing*, pages 1125–1128, Nov 1990.

[45] J.H. Hansen and M. A. Clements. Iterative speech enhancement with spectral constraints. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 189–192, Dallas, April 1987.

[46] B. A. Hanson and T. H. Applebaum. Features for noise-robust speaker-independent word recognition. In *International Conference on Speech and Language Processing*, pages 1117–1120, 1990.

[47] B. A. Hanson and T. H Applebaum. Robust speaker-independent word recognition using static, dynamic and acceleration features: experiments with Lombard and noisy speech. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 857–860, Albuquerque, New Mexico, April 1990.

[48] B. A. Hanson and T. H. Applehaum. Subband or cepstral domain filtering for recognition of Lombard and channel-distorted speech. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 79–82, 1993.

[49] B. A. Hanson and H. Wakita. Spectral slope distance measures with linear prediction analysis for word recognition in noise. *IEEE Trans. Acoust., Speech and Signal Processing*, 35(7):968–973, 1987.

[50] H. Hermansky and N. Morgan. Towards handling the acoustic environment in spoken language processing. In *International Conference on Speech and Language Processing*, volume I, pages 85–88, Banff, Alberta, Canada, October 1992.

[51] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. Compensation for the effect of communication channel in auditory-like analysis of speech (RASTA-PLP). In *Proceedings of European Conference on Speech Technology*, pages 1367–1370, Genova, Italy, September 1991.

[52] H. Hermansky, N. Morgan, and H.-G. Hirsch. Recognition of speech in additive and convolutional noise based on RASTA spectral processing. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 83–86, 1993.

[53] X. D. Huang. A study on speaker-adaptive speech recognition. In *Speech and Natural Language Workshop*. DARPA, February 1991.

[54] M. J. Hunt, D. C. Bateman, S. M. Richardson, and P. Piau. An investigation of PLP and IMELDA acoustic representation and of their potential for combination. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 881–884, Toronto, Canada, May 1991.

[55] M. J. Hunt and C. Lefebvre. A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 1989.

[56] D. R. Hush and B. G. Horne. Progress in supervised neural networks. *IEEE Signal Processing Magazine*, pages 8–39, January 1993.

[57] B. H. Juang and K. K. Paliwal. Vector equilization in hidden Markov models for noisy speech recognition. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 301–304, 1992.

[58] B. H. Juang and L. R. Rabiner. Signal restoration by spectral mapping. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing 1987*, pages 2368–2371, 1987.

[59] B. H. Juang, L. R. Rabiner, and J. G. Wilpon. On the use of bandpass lifting in speech recognition. *IEEE Trans. Acoust., Speech and Signal Processing*, 35(7):947–954, 1987.

[60] J.-C. Junqua. The Lombard reflex and its role on human listeners and automatic speech recognizers. *J. Acoustic. Soc. Am.*, 93(1):510–524, Jan. 1993.

[61] J.-C. Junqua and Y. Anglade. Acoustic and perceptual studies of Lombard speech: application to isolated-words automatic speech recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 841–844, Albuquerque, New Mexico, 1990.

[62] T. Kitamura, S. Ando, and E. Hayahara. Speaker-independent spoken digit recognition in noisy environments using dynamic spectral features and neural networks. In *International Conference on Speech and Language Processing*, volume I, pages 699–702, Banff, Alberta, Canada, October 1992.

[63] D. H. Klatt. A digital filterbank for spectral matching. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 573–576, 1976.

[64] H. Kobatake, K. Gyoutoku, and S. Li. Enhancement of noisy speech by maximum likelihood estimation. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 973–976, Toronto, Canada, May 1991.

[65] C. H. Lee, C. H. Lin, and B. H. Juang. A study on speaker adaptation of the parameters of continuous density hidden Markov models. *IEEE Trans. on Signal Processing*, 39(4):806–814, April 1991.

[66] C. Lefebvre, D. Zwierzynski, D. Starks, and G. Birch. Further optimization of a robust IMELDA speech recogniser for applications with severely degraded speech. In *International Conference on Speech and Language Processing*, volume I, pages 691–694, Banff, Alberta, Canada, October 1992.

[67] J. Lim and A. Oppenheim. All pole modeling of degraded speech. In J. Lim, editor, *Speech Enhancement*, pages 101–114. Prentice-Hall, Englewood Cliffs, NJ, 1983.

[68] J. S. Lim. Spectral root homomorphic deconvolution system. *IEEE Trans. on Acoust., Speech and Signal Processing*, June 1979.

[69] J. S. Lim. *Speech Enhancement.* Prentice-Hall, Englewood Cliffs, NJ, 1983.

[70] J. S. Lim and A. V. Oppenheim. Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition. *IEEE Trans. on Acoust., Speech and Signal Processing*, 26(4):354–358, 1978.

[71] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for the vector quantizer design. *IEEE. Trans. on Communication*, COM-28(1):84–95, Jan. 1980.

[72] R. P. Lippmann. An introduction to computing with neural nets. *IEEE ASSP Magazine*, 3:422, April 1987.

[73] P. Lockwood, C. Baillargeat, J. M. Gillot, J. Boudy, and G. Faucon. Noise reduction for speech enhancement in cars: Non-linear spectral subtraction – Kalman filtering. In *Proceedings of EuroSpeech*, 1991.

[74] P. Lockwood and J. Boudy. Experiments with a nonlinear spectral substractor (nss), hidden markov models and the projection, for robust speech recognition in cars. *Speech Communication*, 11:215–228, 1992.

[75] D. Mansour and B. H Juang. A family of distortion measures based upon projection operation for robust speech recognition. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 36–39, New York, April 1988.

[76] D. Mansour and B. H Juang. The short-time modified coherence representation and its application for noisy speech recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 525–528, New York, April 1988.

[77] D. Mansour and B. H. Juang. A family of distorsion measures based upon projection operation for robust speech recognition. *IEEE Trans. on Acoust., Speech and Signal Processing*, ASSP-37(11), 1989.

[78] J. D. Markel and A. H. Gray Jr. *Linear Prediction of Speech.* Springer-Verlag, New York, 1976.

[79] T. Matsuoka and K. Shikano. Speaker adaptation by modifying mixture coefficients of speaker-independent mixture Gaussian HMMs. In *International Conference on Speech and Language Processing*, volume I, pages 373–376, Banff, Alberta, Canada, October 1992.

[80] B. A. Mellor and A. P. Varga. Noise masking in a transform domain. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 87–90, 1993.

[81] J. G. Mena, L. S. Sandoval, and R. G. Gomez. A comparative study of feature extraction methods for noisy speech recognition. In L. Torres, E. Masgrau, and M. A. Lagunas, editors, *Signal Processing V: Theories and Applications*, pages 1191–1194. Elsevier Science Publishers B. V., 1990.

[82] S. Mizuta and K. Nakajima. Optimal discriminative training for HMMs to recognize noisy speech. In *International Conference on Speech and Language Processing*, volume II, pages 1519–1522, Banff, Alberta, Canada, October 1992.

[83] C. Mokbel, L. Barbier, Y. Kerlou, and G. Chollet. Word recognition in the car: Adapting recognisers to new environments. In *International Conference on Speech and Language Processing*, volume I, pages 707–710, Banff, Alberta, Canada, October 1992.

[84] C. Mokbel and G. Chollet. Speech recognition in adverse environments: speech enhancement and spectral transformations. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 925–928, 1991.

[85] S. Morii, T. Morii, and M. Hoshimi. Noise robustness in speaker independent speech recognition. In *International Conference on Speech and Language Processing*, pages 1145–1148, Nov 1990.

[86] A. Nadas, D. Nahamoo, and M. A Picheny. Speech recognition using noise-adaptive prototypes. *IEEE trans. on ASSP*, 37(10):1495–1502, October 1989.

[87] S. Nakamura and K. Shikano. Speaker adaptation applied to HMM and neural networks. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 89–92, 1989.

[88] S. Nakamura and K. Shikano. A comparative study of spectral mapping for speaker adaptation. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 157–160, Albuquerque, New Mexico, April 1990.

[89] N. Nocerino, F. K. Soong, L. R. Rabiner, and D. H. Klatt. Comparative study of several distorsion measures for speech recognition. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, pages 25–28, 1985.

[90] K. Ohkura, D. Rainton, and M. Sugiyama. Noise-robust HMMs based on minimum error classification. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 75–78, 1993.

[91] K. Ohkura, M. Sugiyama, and S. Sagayama. Speaker adaptation based on transfer vector field smoothing with continuous mixture density HMMs. In *International Conference on Speech and Language Processing*, volume I, pages 369–372, Banff, Alberta, Canada, October 1992.

[92] D. O'Shaughnessy. Speech enhancement using vector quantization and a formant distance measure. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 549–552, 1988.

[93] D. O'Shaughnessy. Enhancing speech degraded by additive noise or interfering speakers. *IEEE Communications Magazine*, pages 46–52, Feb. 1989.

[94] K. K. Paliwal. Neural net classifier for robust speech recognition under noisy environments. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 429–432, Albuquerque, New Mexico, April 1990.

[95] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill Book Company, second edition, 1984.

[96] J. E. Porter and S. F. Boll. Optimal estimators for spectral restoration of noisy speech. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 18A2.1–2.4, 1984.

[97] T. F. Quatieri and R. J. McAulay. Noise reduction using a soft-decision sine-wave vector quantization. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 821–824, 1990.

[98] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, February 1989.

[99] R. Roth, J. Baker, J. Baker, L. Gillic, M. Hunt, Y. Ito, S. Lowe, J. Orloff, B. Peskin, and F. Scattone. Large vocabulary continuous speech recognition of Wall street journal data. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 640–643, 1993.

[100] S. A. Shamma. Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. *J. Acoust. Soc. Amer.*, pages 1622–1632, 1985.

[101] K. Shikano, K. F. Lee, and R. Reddy. Speaker adaptation through Vector Quantization. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Tokyo, 1986.

[102] O. Siohan, Y. Gong, and J.-P. Haton. A Bayesian approach to phone duration adaptation for Lombard speech recognition. In *Proceedings of European Conference on Speech Communication and Technology*, Berlin, Germany, September 1993.

[103] F. K. Soong and M. M. Sondhi. A frequency-weighted Itakura spectral distortion measure and its application to speech recognition in noise. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 625–628, Dallas, April 1987.

[104] H. W. Sorenson. *Parameter estimation: Principles and problems.* Marcel Dekker, NY, 1980.

[105] R. M. Stern and A. Acero. Acoustical pre-processor for robust speech recognition. In *Proc. DARPA Workshop on Speech and Natural Language*, pages 311–318, 1990.

[106] R. M. Stern and M. J. Lasry. Dynamic speaker adaptation for feature-based isolated word recognition. *IEEE Trans. on Acoust., Speech and Signal Processing*, 35(6), June 1987.

[107] S. Tamura and A. Waibel. Noise reduction using connectionist models. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 553–5562, 1988.

[108] H. P. Tseng, M. J. Sabin, and E. A. Lee. Fuzzy vector quantization applied to hidden Markov modeling. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing 1987*, pages 641–644, Dallas, Texas, April 1987.

[109] A. P. Varga and K. M. Ponting. Control experiments on noise compensation in hidden Markov model based continuous word recognizers. In *Proceedings of European Conference on Speech Technology*, Paris, 1989.

[110] A. P. Verga and R. K. Moore. Hidden Markov model decomposition of speech and noise. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 845–848, 1990.

[111] K. Wang, S. A. Shamma, and W. J. Byrne. Noise robustness in the auditory representation of speech signals. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 335–338, 1993.

Communications
Research Center

Shirleys Bay
3701 Carling Avenue
P.O. Box 11490
Station H
Ottawa, Ontario
K2H 8S2

Centre de recherches
sur les communications

Shirleys Bay
3701, avenue Carling
Case postale 11490
Succursale H
Ottawa (Ontario)
K2H 8S2