Speech Communication 11 (1992) 215–228
North-Holland

Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars

P. Lockwood and J. Boudy
Matra Communication, rue J.P. Timbaud, 78392 Bois d'Arcy Cedex, France

Received 26 September 1991
Revised 23 January 1992

Abstract. Achieving reliable performance for a speech recogniser is an important challenge, especially in the context of mobile telephony applications where the user can access telephone functions through voice. The breakthrough of such a technology is appealing, since the driver can concentrate completely and safely on his task while composing and conversing in a "full" hands-free mode. This paper addresses the problem of speaker-dependent discrete utterance recognition in noise. Special reference is made to the mismatch effects due to the fact that training and testing are made in different environments. A novel technique for noise compensation is proposed: nonlinear spectral subtraction (NSS). Robust variance estimates and robust pdf evaluations (projection) are also introduced and combined with NSS into the HMM framework. We show that the lower limit of applicability of the projection (low SNR values) can be loosened after combination with NSS. Experimental results are reported. The performance of an HMM-based recogniser rises from 56% (no compensation) to 98% after speech enhancement. More than 3300 utterances have been used to evaluate the systems (three databases, two European languages). This result is achieved by the use of robust training/recognition schemes and by preprocessing the noisy speech by NSS.

Zusammenfassung. Leistungsfähige Spracherkenner zu entwickeln ist eine wichtige Forschungsaufgabe. Dies gilt insbesondere auch im Bereich des Mobilfunks, wenn der Benutzer sein mobiles Telefon durch akustische Eingabe bedienen können soll. Derartige Verfahren können beispielsweise dann attraktiv sein, wenn sich ein Autofahrer in die Lage versetzt sieht, Telefonverbindungen zu wählen und Telefongespräche zu führen, ohne seine Hände vom Steuer nehmen zu müssen, und sich somit vollständig und sicher aufs Fahren konzentrieren kann. Der vorliegende Beitrag befaßt sich mit dem Problem sprecherabhängiger Erkennung isolierter Äußerungen in geräuschvoller Umgebung. Hierbei wird insbesondere das Problem diskutiert, das dadurch entsteht, daß die Umgebungsbedingungen beim Training und beim Einsatz des Algorithmus erheblich voneinander abweichen. Präsentiert wird das Verfahren der nichtlinearen spektralen Subtraktion (NSS), eine neuartiges Methode zur Geräuschreduktion. Darüber hinaus werden robuste Schätzverfahren für Varianzen und robuste Evaluierungsverfahren für Wahrscheinlichkeitsdichtefunktionen (Projektionen) eingesetzt und zusammen mit dem NSS-Verfahren in ein Spracherkennungssystem auf HMM-Basis eingebaut. Wie gezeigt wird, kann der minimale Störabstand, bei dem der beschriebene HMM-Erkenner noch funktioniert, durch den Einsatz des NSS-Verfahrens erheblich gesenkt werden. Experimentelle Ergebnisse werden vorgestellt. Die Erkennungsrate des HMM-Spracherkenners wächst von 56% (ohne Geräuschkompensation) auf 98% (mit Einsatz aller beschriebenen Verfahren). Zur Evaluierung des Systems wurden mehr als 3300 Äußerungen verwendet (drei Korpora, zwei europäische Sprachen). Die Verbesserung wurde erzielt durch den Einsatz robuster Verfahren in der Lern- und Betriebsphase des Erkenners sowie durch Qualitätsverbesserung des gestörten Sprachsignals mit dem NSS-Verfahren.

Résumé. Atteindre des performances robustes pour un système de reconnaissance vocale est un problème difficile à résoudre surtout lorsqu'un tel système est utilisé comme fonction de composition vocale dans les radiotéléphones mobiles de voiture. La nécessité de telles fonctions devient primordiale dans la mesure où l'utilisateur d'un radiotéléphone mobile peut se concentrer sans risques sur la conduite de son véhicule tout en composant le numéro de son correspondant et discuter avec ce dernier en mode "mains-libres". Le travail présenté dans cet article pose le problème de la reconnaissance mono-locuteur de mots isolés dans un environnement bruité. Dans ce contexte toute la difficulté réside dans le fait qu'il existe des différences importantes entre les conditions d'apprentissage (généralement dans le silence) et celles de reconnaissance (généralement dans le bruit, lorsque le véhicule roule). Une nouvelle technique de réduction du bruit est proposée: la Soustraction Spectrale Non linéaire (NSS). Dans un système de reconnaissance utilisant les Modèles de Markov Cachés (HMM), des estimateurs robustes de variances (lissage) et de densités de probabilités d'observation (projection) sont également introduits et combinés avec la Soustraction Spectrale Non linéaire. Nous montrons aussi que les limites courantes d'application de la Projection (RSB

inférieurs à 0 dB) peuvent être repoussées grâce à l'utilisation de NSS. Des simulations numériques faites à partir de données réelles sont présentées et commentées. Le système de reconnaissance (HMM) voit ses performances s'élever de 56%, sans traitement, à 98%, après réduction du bruit par NSS. Plus de 3000 mots à reconnaître ont été employés pour l'évaluation des différents systèmes considérés (trois bases de données, deux langues européennes). De telles performances ont été atteintes en ayant recours à des techniques robustes d'apprentissage et de reconnaissance ainsi qu'à un prétraitement des mots bruités à l'aide de NSS.

Keywords. Speech recognition; projection measure; speech enhancement; spectral subtraction; noise; continuous density Hidden Markov Model.

1. Introduction

The problem investigated here is that of speech recognition in noisy environments. The use of speech recognition in the car will allow flexibility and increase security aspects as the driver can concentrate on his driving task and access some commands by voice. While some functions in the car can correspond to an a priori determined vocabulary (examples: commands for the radio, the air conditioning, digits and numbers for dialling) others require speaker-tuned vocabularies.

Voice access to an agenda by uttering a single word (typically the name of the desired subscriber) will avoid a time consuming and cumbersome dialling phase. Dialling necessitates reliable performance at the string level which implies a very high recognition score at the utterance level. While for fixed vocabularies a multi-speaker multi-conditions approach could be envisaged (learning in noise), in an agenda-driven access mode a speaker dependent approach must be used. In this case the system has to deal with the mismatch problem, in the sense that training is made in a quiet and unstressful environment (car stopped), while the recognition is made in completely different conditions: we impose training with car stopped for safety reasons. The mismatch is not only due to noise that will corrupt the test signal, but also to the increased variability of pronounciation caused by various factors such as stress, tiredness or simply the noise level (Lombard effect).

In the following, we report experiments made on two types of well known recognisers: a DTW-based system, and a continuous density HMM-based system.

This latter type of system faces another important problem: HMMs do not perform well if the parameters of the models are underestimated, and this is caused by lack of sufficient training material. At least 20 or more repetitions of each vocabulary word are needed to have a reasonable estimation of the HMM parameters. In a speaker-dependent context, it is simply not acceptable to ask a user to pronounce so many repetitions, even for a small vocabulary. The right figure could probably be between 2 and 4 repetitions maximum. In this case, robust HMM training/recognition schemes are necessary.

The problem of automatic speech recognition in noise has received much interest in the scientific community. A good review has been presented in a recent publication (Juang, 1991). For signal enhancement, we propose a novel nonlinear spectral subtraction technique (NSS). This method makes use of speech and noise knowledge, is applied at frame level, and can be considered as combining noise subtraction and noise masking properties in the same framework. A robust pdf evaluation is also introduced with the use of the projection measure extended to HMMs (Carlson and Clement, 1991). We have extended further the technique and we put forward two new results:

- A modified projection performs better than the "standard" projection.
- The results with the projection measure are significantly improved when combined with an appropriate noise compensation technique such as NSS.

The next section reviews the nonlinear subtractor. Section 3 presents a robust HMM-based recogniser. Two aspects of the robustness are reported and refer to robust variance estimates and robust-to-noise pdf evaluations. This resulted in the development of a modified Viterbi scheme making use of the cepstral projection technique, the result being an algorithm that performs well in noise without any necessary preprocessing of the noisy

speech. Section 4 describes the context of experiments. Section 5 describes preliminary experiments made with a DTW-based system. Section 6 describes results obtained with the various HMM recognisers. Section 7 concludes this contribution.

2. Nonlinear spectral subtraction

2.1. General considerations

Car noise

Some experiments have been carried out in several vehicles. To analyse the noise, recordings were made in various conditions of noise: several speeds, several environments (urban areas, highways, open country roads). The result we found was that noise was incoherent when comparing recordings made simultaneously using several microphones put in different places in the car (Lecomte et al., 1989). The noise model can be considered as a combination of several noise sources caused by engine, tyres (containing basically low frequency components) and aerodynamic turbulences having a broader spectrum.

Choice of a sensor

The sensor is the first element of the input chain, therefore selecting a "good" sensor will extend the applicability of the techniques. This problem has been studied. It came out that the best microphone was a head-mounted noise-cancelling microphone. Unfortunately, such a sensor is not desirable in our application of interest as a "hands-free" solution is preferred. Omni-, bi- and uni-directional microphones were also compared. Uni-directional microphones clearly outperformed the others. Eleven microphones/seven positions were compared, based on measurements such as the signal-to-noise ratio, a distortion measure and a speech recognition score. A methodology exhibiting low variability was introduced, mainly based on the use of an artificial mouth so as to reproduce source invariance. This source was then played in different conditions of noise. Two microphones finally came out as giving the best overall results (Boudy et al., 1990).

Mono/multi sensor input schemes

These studies focus on mono-sensor solutions for several reasons:

- Industrial constraint is the first aspect as the overhead cost induced by multi-sensor schemes can be a barrier, at least for a low-cost product.
- We felt that measuring the limits of applicability of one-sensor schemes was an important step before considering more complex structures.

Mono-sensor schemes have several drawbacks though. From a state of the art analysis performed previously (Baillargeat et al., 1990), it was shown that methods that do not require explicit segmentation constraints do not perform well. Therefore the speech/noise segmentation process is an important feature of one-sensor based systems. This segmentation operation could be difficult at low SNRs. Another drawback is that the noise estimates that are computed during speech pauses are not re-estimated during speech. This could cause problems if noise varies rapidly during speech. In these cases, combining mono- and multi-sensor solutions could be an attractive way to tackle the problem.

2.2. Nonlinear spectral subtraction

Spectral subtraction techniques have proved efficient for speech recognition in noise (Compernolle, 1989). The basics of this new nonlinear spectral subtraction technique (NSS) resides in the combination of two main ideas:

- The use of an extended noise model composed of an averaged noise vector and an overestimation model, estimated jointly during the speech pauses.
- A nonlinear implementation of the subtraction process, taking into account the frequency-dependent signal to noise ratio (SNR), the idea being to apply a minimal subtraction factor in high SNR regions at frame level.

In the standard spectral subtraction method (referred to in the following as linear spectral subtraction (Berouti et al., 1979; Boll, 1979; Paul, 1980; Lockwood et al., 1991)), the short term spectral magnitude of noise is subtracted to the noisy speech as follows:

$$Y_i[\omega] = H_i[\omega]X_i[\omega],$$

where

$$H_i[\omega] = \frac{D_i[\omega]}{|\ddot{X}_i[\omega]|}$$

and

$$D_i[\omega] = |\ddot{X}_i[\omega]| - |\ddot{B}_i[\omega]|,$$

with

$X_i[\omega]$: short term estimation of speech (frame i),
$B_i[\omega]$: short term estimation of noise (frame i),
$|\ddot{X}_i[\omega]|$: smoothed estimate of the corrupted speech magnitude at time i,
$|\ddot{B}_i[\omega]|$: smoothed estimate of the noise magnitude at time i,
$Y_i[\omega]$: clean speech estimate

$|\ddot{X}_i[\omega]|$ and $|\ddot{B}_i[\omega]|$ are obtained using the following forgetting factors λ_X and λ_B:

$$|\ddot{B}_i[\omega]| = \lambda_B |\ddot{B}_{i-1}[\omega]| + (1-\lambda_B)|B_i[\omega]|,$$

$$|\ddot{X}_i[\omega]| = \lambda_X |\ddot{X}_{i-1}[\omega]| + (1-\lambda_X)|X_i[\omega]|.$$

Generally, we have

$$0.1 \leqslant \lambda_X \leqslant 0.5,$$

$$0.5 \leqslant \lambda_B \leqslant 0.9.$$

In the nonlinear spectral subtraction scheme, an improved noise model is determined: $(|\ddot{B}[\omega]|, \alpha(\omega))$. $\alpha(\omega)$ is a frequency-dependent overestimation factor that can be estimated during the speech pauses jointly with the noise magnitude. $\alpha(\omega)$ is computed over the last frames of noise so that at time i (note that we have a 50% frame overlap)

$$\alpha_i(\omega) = \max_{i-40 \leqslant \tau \leqslant i} (|B_\tau[\omega]|).$$

Having an estimation of both the noise magnitude $|\ddot{B}_i[\omega]|$ and the noise overestimation model $\alpha_i(\omega)$, the subtraction can be implemented nonlinearly as follows:

$$D_i[\omega] = |\ddot{X}_i[\omega]| - \Phi(\rho_i[\omega], \alpha_i(\omega), |\ddot{B}_i[\omega]|),$$

with

$$\rho_i[\omega] = \frac{|\ddot{X}_i[\omega]|}{|\ddot{B}_i[\omega]|},$$

$\rho_i[\omega]$ is a biased estimate of the signal to noise ratio at frame i.

To prevent $D_i[\omega]$ from taking negative values as well as to introduce a noise-masking effect, a flooring is applied (see below).

Φ is a nonlinear function, weighting the subtraction process according to the signal to noise ratio in a specific frequency band with upper and lower limits:

$$|\ddot{B}_i[\omega]| \leqslant \Phi(\rho_i[\omega], \alpha_i(\omega), |\ddot{B}_i[\omega]|) \leqslant 3|\ddot{B}_i[\omega]|.$$

$\Phi(\omega)$ can be chosen as any arbitrary function implementing the following idea: apply a minimum subtraction factor in high SNR regions and subtract more noise in low SNR regions, until a "noise masking" threshold is reached. The following function has been used in these experiments:

$$\Phi(\rho_i[\omega], \alpha_i(\omega), |\ddot{B}_i[\omega]|) = \frac{\alpha_i(\omega)}{1+\gamma\rho_i[\omega]}.$$

γ is a scaling factor depending on the variation range of the frequential SNR $\rho_i[\omega]$.

Optimisation of the nonlinear function

Other functions have been tested and compared, based on the following decomposition:

$$\Phi(\rho_i[\omega], \alpha_i(\omega), |\ddot{B}_i[\omega]|) = \alpha_i(\omega) F_n(\rho_i[\omega], \alpha_i(\omega), |\ddot{B}_i[\omega]|),$$

where F_n denotes various NSS schemes (see Table 1). The figures in Table 1 correspond to tests made on the MATRA database after downsampling to 8 kHz (Lockwood et al., 1992a) and give an idea on the relative behaviour of each function.

We see that when the SNR can be evaluated with sufficient accuracy (e.g., noise conditions recorded at 90 km/h), NSS can still be significantly optimised. F_1 corresponds to a linear subtraction. F_2–F_4 are nonlinear schemes. The basic difference between F_2–F_4 resides in how we are able to make full use of the nonlinear aspect of NSS, with respect to the SNR. Basically, F_4 implements more efficiently the main idea of NSS which is to apply less subtraction effect in higher SNR regions. "lin" refers to a linear weighting function of the SNR, while "sig" denotes the implementation of a sigmoid function. $\rho_{\min}$ and $\rho_{\max}$ are minimum and maximum estimates of the SNR, using a robust estimator (see below).

Table 1
Set of functions used in the NSS schemes

	F_1	F_2	F_3	F_4
$\alpha_i(\omega)$	$\lvert\ddot{B}_i[\omega]\rvert$	$\max_\tau \lvert B_\tau[\omega]\rvert$	$\max_\tau \lvert B_\tau[\omega]\rvert$	$\max_\tau \lvert B_\tau[\omega]\rvert$
F_n	1	1	$1-\left(1-\frac{\lvert\ddot{B}_i[\omega]\rvert}{\alpha_i(\omega)}\right)\mathrm{lin}(\omega)$	$1-\left(1-\frac{\lvert\ddot{B}_i[\omega]\rvert}{\alpha_i(\omega)}\right)\mathrm{sig}(\omega)$,
			$\mathrm{lin}(\omega)=f(\rho_{\min}, \rho_{\max})$	$\mathrm{sig}(\omega)=g(\rho_{\min}, \rho_{\max})$
Improvement 90 km/h		+6%	+8%	+9.5%
Improvement 130 km/h		+15.5%	+16%	+16%

Noise masking and spectral flooring

The spectral flooring is an important factor that will prevent both the resulting subtraction to become negative as well as leave the masked speech untouched. The flooring is implemented as follows (Berouti et al., 1979):

$$D_i[\omega]=\begin{cases}D_i[\omega] & \text{if } D_i[\omega]\geq\beta|\ddot{\boldsymbol{B}}_i[\omega]|,\\ \beta|\ddot{\boldsymbol{X}}_i[\omega]|.\end{cases}$$

A typical value for β is 0.1.

Estimation of the overestimation model

Berouti et al. have used an overestimation of the noise model as a mean to remove the so-called "musical tones" inherent to the spectral subtraction techniques when applied to speech transmission problems (Berouti et al., 1979; Lockwood et al., 1991). The overestimation model proposed by Berouti is constant and has to be defined "a priori". Faucon et al. (1991) have shown that a "good" compromise value is $\alpha_i(\omega)=1.5|\ddot{\boldsymbol{B}}_i[\omega]|$, for all i. In this case, the overestimation is a constant function of the noise model. We have observed that the optimal value for this weighting factor was dependent on the condition of noise. Furthermore, this overestimation was also found to be frequency dependent, as the variance of the estimator is proportional to the noise dynamics and therefore much more important in the low frequencies (for car noise). With NSS, the frequency-dependent overestimator is estimated during speech pauses jointly with the noise model. The estimation of this model $\alpha_i(\omega)$ is computationally complex, though it can be significantly simplified (Lockwood and Boudy, 1992b).

Evaluation of the SNR

The evaluation of the SNR in noise when only a single corrupted channel is available is a problem in itself (Ruehl et al., 1991). The best results have been obtained by evaluating the SNR using a separate smoothed estimate of the short term speech, having a stronger smoothing factor than the one used in the subtraction process (typical values are $\lambda_X=0.1$ for use in the subtraction, $\lambda_X=0.5$ for use in the SNR evaluation).

3. Robust Hidden Markov Models

The recognition system is based on a continuous density HMM recogniser with an optimised topology (16 states, no skip). For each state in the model, the distribution is represented by a unimodal Gaussian density or Modified Gaussian density as we shall see in the following.

When insufficient training material is available, it is necessary to use robust estimators (Paul et al., 1986). The fixed variance estimate is one of such estimators. A smoothed variance estimate is possible to use also.

Modified Gaussian density functions: the projection and modified projection measures

The projection measures (Paliwal, 1982; Mansour and Juang, 1989) have proved to be efficient in a DTW-based system; therefore the extension to HMMs was appealing. Even though such work has been reported by Carlson and Clements (1991) recently, our contribution (developed independently) has further extended the technique as we show that the projection measure is really

efficient in HMMs with car-noise when welded with an appropriate noise compensation technique (NSS). On the other hand we also compared the projection measure with the modified projection measure (Mansour and Juang, 1989). Both have been extended to HMMs and compared to the standard (Euclidean/Gaussian) case. This comparison has been performed with and without noise compensation.

The continuous observation probability density functions (pdf) for an N-states HMM are given by the following expression:

$$b_j(O) = \sum_m c_{jm} N_{qk}\{O, \mu_{jm}, \Sigma_{jm}\}$$

for the state j with $1 \leqslant j \leqslant N$,

where c_{jm} denotes the mixture coefficient for the mth mixture component $N_{qk}\{\cdot\}$ (case of M components). Note that in this work, the variance–covariance matrices are taken diagonal, assuming independence between the components of a feature vector.

The general formulation of the Gaussian density N_{qk} can be written as

$$N_{qk}\{O, \mu_{jm}, \Sigma_{jm}\} = [(2\Pi)^p |\Sigma_{jm}|]^{-1/2} \exp(-0.5 d(q, k)),$$

with

$$d(q, k) = [(O - q\mu_{jm})^T \Sigma_{jm}^{-1} (O - q\mu_{jm})]^k, \tag{3.1}$$

where q can correspond to the "under-weighting" factor.

(a) With $q = 1$ and $k = 1$, we have the standard Gaussian pdf (referred to in the following as state-variances).

(b) With $q = 1$, $k = 1$, $\Sigma_{jm} = \Sigma$, for all j and m, we have the fixed variance case. Σ is computed from the available training material for one speaker (Σ is computed speaker-dependently).

(c) With $q = 1$, $k = 1$, $\Sigma_{jm} = W^{-1}$, we have the smoothed variance. The coefficients of W are those of an appropriate smoothing function as discussed below (see robust variance estimates).

(d) With an optimal value of q (q^*) and $k = 1$, $\Sigma_{jm} = I$, we have the projection measure (Mansour and Juang, 1989). The "under weighting" factor q^* is determined in order to minimise the expression (3.1):

$$q^* = \frac{O^T \mu_{jm}}{|\mu_{jm}|^2}. \tag{3.2}$$

By reporting (3.2) in (3.1), we obtain a new formulation of the density function. The expression for the projection measure is then

$$d(q^*, 1) = |O|^2 (1 - \cos^2 \beta),$$

where

$$\cos \beta = \frac{O^T \mu_{jm}}{|O|\,|\mu_{jm}|},$$

β is the angle between the noisy cepstral vector O and the noiseless mean vector μ_{jm}.

(e) With $k = \frac{1}{2}$, $q = q^*$, $\Sigma_{jm} = I$, we have the modified projection (Mansour and Juang, 1989). In this latter case, if we suppose that the angle between the vectors is small, we have

$$d'(O, \mu_{jm}) = |O|(1 - \cos \beta),$$

where d' denotes the modified projection.

By deriving (3.3) depending on β, we have

$$d'(O, \mu_{jm}) = |O|(2(\sin^2(\beta/2)).$$

So, by assuming β small (less than 0.5 radius) which is realistic as shown in (Mansour and Juang, 1989), we have

$$d'(O, \mu_{jm}) = |O| \frac{\beta^2}{2}. \tag{3.4}$$

On the other hand, $d(q^*, 1)$ can be approximated, for small values of β as follows:

$$d(q^*, 1) = |O|^2 (\sin^2 \beta) = (|O|\,|\beta|)^2;$$

so by taking the square root value we have

$$d(q^*, 0.5) = |O|\,|\beta|. \tag{3.5}$$

So the only difference between (3.4) and (3.5) is the value of the exponent of β. So, for small values of β we have

$$d'(O, \mu_{jm}) < d(q^*, 0.5).$$

This shows that (3.4) is a minimised version of the square root of $d(q^*, 1)$. Anyhow, these two measures are of the same kind. So the normal densities of the form $A \exp(-0.5x^2)$ are "modified"

with the shape of the Laplacian form: $A \exp(-0.5|x|)$.

(f) With $q=q^*$, $k=1$, $\Sigma_{jm}=W^{-1}$ or $q=q^*$, $k=0.5$, $\Sigma_{jm}=W^{-1}$ taking any form defined below, we have the weighted projection and the weighted modified projection (smoothed variance case).

Robust variance estimates: fixed/smoothed variances

A weighted Euclidean measure is used implicitly in continuous density (Gaussian) Hidden Markov Models. With the hypothesis of equi-covariance ($\Sigma_{jm}=\Sigma$) for all states, the Euclidean is weighted by a fixed variance (Paul, 1986). Other weightings can be introduced, replacing the variances by an appropriate smoother. Band-pass lifters (Juang et al., 1987), RPS (Tokhura, 1987), exponential (Hermansky and Junqua, 1988) have been used efficiently. These functions have been experimented and compared.

Figure 1 plots several smoothing functions computed on MATRA database, for one speaker (the behaviour is similar with other speakers). In Figure 1(a), the fixed variance is drawn together with its polynomial approximation and the optimal smoother. The optimal smoother corresponds to the function yielding the best results as shown in Figure 3 (exponential lifter, $\alpha=0.35$). In Figure 1(b), a "best fit" smoother is the smoothing function closest to the fixed variance (exponential lifter, $\alpha=0.6$). The following behaviour has been observed: the polynomial approximation improves the results obtained with the fixed variance but the optimal smoother outperforms both representations. The optimal smoother also gives better results than the "best fit" smoother (Figure 3).

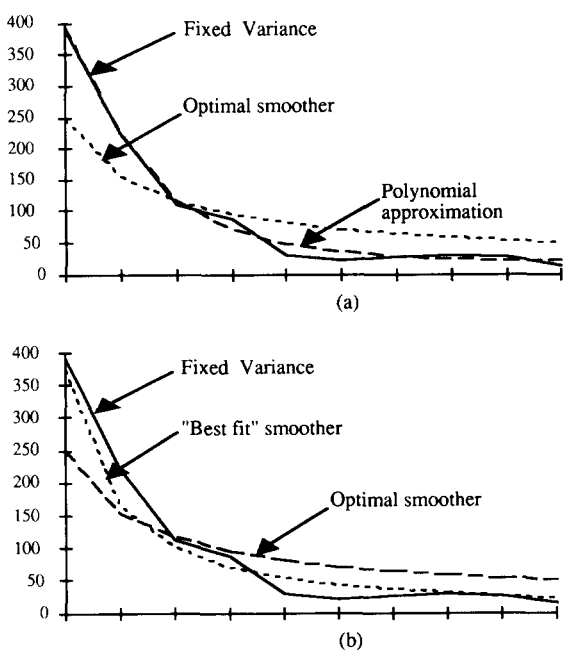


Fig. 1. Comparison of the various smoothers including a least-squares approximation of the fixed variance by a polynome of degree 3. The fixed variance has been computed for one speaker of the database MATRA (4 repetitions recorded in silence).

We demonstrate a significant improvement in performance using appropriate weightings integrated in either the Euclidean, the projection or the modified projection measures. These modifications are introduced in the Viterbi algorithm during the recognition phase. The projection has not been used during training as it was done with car stopped and therefore in "clean" conditions; though the projection could also efficiently be used to robustise a training in noise (Carlson and Clements, 1991).

Scaling of the variance smoothers

A scaling factor is used with the variance smoothers. The scaling is computed in order to assure comparable variations between the fixed variance and a particular smoother. The optimal "gain" factor γ is obtained by minimising the following criterion:

$$J(\gamma)=\sum_{i=1}^{p}\left(\frac{\Sigma_i-\gamma W_i^{-1}}{\Sigma_i}\right)^2,$$

Σ_i designs the ith component of the fixed variance (diagonal term), W_i^{-1} the inverse of the ith component of a particular smoother and p is the number of cepstral parameters. γ is applied to any smoothing function except for the modified weighted projection (in this case $\gamma=1$).

4. The ARS databases

The evaluation has been made on databases recorded within the ESPRIT-ARS project by several partners of this project. In this case the databases recorded by ENST, MATRA and CSELT have been used. Table 2 summarises the

Table 2
Description of the database used during training and testing

Database	Number of speakers	Number of words	Number of utterances for training	Number of utterances for testing recorded at: 70 km/h	90 km/h	110 km/h	130 km/h
MATRA	4	43	4		4		4
ENST	1	43	10			15	
CSELT	4	34	10	5			5

characteristics of these databases. Each speaker pronounced the set of utterances.

In all these tests, the frame rate has been set at 16 ms with a window width of 32 ms. The number of cepstral parameters used was set to 10. The sampling frequency was fixed at 16 kHz.

In the first phase of the evaluation (Section 5), the methods are tested on MATRA's database. These concern the results obtained with the DTW-based system. In this case an automatic end-pointing was used but the gross segmentation errors were rejected.

In the second phase (tests of the HMMs – Section 6) the evaluation has been made on several databases (ENST, MATRA and CSELT) in order to appreciate the robustness of our optimised HMM recogniser with different conditions of training and different noise conditions. No automatic segmentor was used. The MATRA and CSELT databases were recorded with a unidirectional microphone placed on the side of the windscreen while the ENST database was recorded with an omnidirectional microphone (catching more ambient noise than a unidirectional microphone) placed relatively far from the speaker. This allows us to dispose of various conditions of car noise varying from reasonable (CSELT 70 km/h, MATRA 90 km/h) to very difficult noise conditions (ENST).

The utterances recorded in silence (car stopped) are used for training the system, while those recorded in noise are used for testing. It should be emphasised that the tests are made on signals recorded in context and not with a posteriori addition of noise.

5. Initial experiments with a DTW-based system

In these first experiments, we studied how a standard DTW scheme could cope with noise and especially the mismatch effect. We compared speech representations and robust distances (lifters, projection). Then, an extension of the DTW results to the HMMs has been investigated.

Comparison of two speech representations

Two kinds of representations have been experimented:

- a parametric representation based on the LPC analysis. LPCC parameters have been used (Linear Prediction Cepstrum Coefficients),
- a non-parametric representation based on Mel-Frequency subbands decomposition of the speech signal, i.e. the MFCC representation (Mel-Frequency Cepstrum Coefficients: (Davis and Mermelstein, 1980)).

Noise compensation techniques have also been welded into the standard LPCC and MFCC computations in order to reduce the mismatch produced by the noise contributions during the recognition process. From first evaluations (Baillargeat et al., 1990) two noise reduction techniques were retained for further investigation; Kalman filtering and Spectral subtraction. These methods have been extended yielding new algorithms for car noise processing. The comparison of the two approaches is reported in (Lockwood et al., 1991). We observed that MFCC parameters performed better than LPCCs if used with an appropriate noise compensation technique.

Comparison of some robust distance weightings

Some noise-robust distance weighting techniques have recently been introduced for reducing the mismatch between noisy signals and clean reference signals. Bandpass (Juang et al., 1987), exponential lifters (Tohkura, 1987; Hermansky and Junqua, 1988) and projection distances (Mansour and

Juang, 1989; Paliwal, 1982) have been experimented in the DTW-based framework and they have shown significant improvements especially when the mismatch between the training and testing conditions was important.

It appears that the NSS technique welded into the MFCC representation, and combined with an appropriate distance weighting gives the best recognition scores. On the other hand, in great mismatch conditions (database MATRA 130, no noise compensation), the projection measure with a well-tuned weighting outperforms the weighted Euclidean distances.

Extension of DTW results to HMMs

The HMM is described in Section 3. Some results are reported here comparing both systems. For the homogeneity of the results, tests with the HMM were computed in the same conditions as the DTW scheme (Section 4 – first phase).

Table 3 summarises the results obtained in the DTW and HMM contexts respectively. It appears that some results are extendable to HMM while others are not. Namely, when the mismatch between training and testing conditions is important, DTW systems outperform HMMs, while on the other hand, when an efficient noise compensation is applied (NSS), the results are reversed and HMMs perform better. One possible explanation is that DTW applies strong temporal constraints while HMMs do not; in other words, the duration model is better represented in the DTW model rather than the HMM one. One alternative could be to use "topology adapted" HMM models (Picone, 1989; Lockwood et al., 1992a).

6. Results with Hidden Markov Models

In the following, results are given combining the various points developed previously, according to the methodology defined in Section 4, phase 2. In the tables, information giving an idea of the variability of the results is given: standard deviation computed across all speakers, noise conditions and databases.

REMARK. The figures presented in Tables 4–6 are slightly modified compared to those initially presented. The difference concerns only the variance smoother, as in the Eurospeech presentation, an "optimal" smoother was selected as a function of the noise condition, while in these tables we have used a "compromise" smoother, constant across all databases and noise conditions. Figures 2 and 3 anyhow contain the detailed results for all smoothers.

State variances/fixed variance/smoothed variance estimate

Table 4 summarises the results obtained with the fixed and smoothed variance estimates for each database tested. The term "Normal" in the table means no noise compensation.

On average, the smoothed variance performs better than the fixed variance estimate.

The good results obtained with the database of CSELT without any noise preprocessing are due to the fact that the training utterances contain noise (engine on, recordings made in a noisy car environment); this reduces the mismatch effect.

Nonlinear spectral subtraction constantly improves the results.

Table 3
Comparative results DTW/HMM using compensation (NSS)

Method / Database	DTW weighted Euclidean (no noise compensation)	HMM smoothed variance (no noise compensation)	DTW weighted Euclidean (NSS)	HMM smoothed variance (NSS)
MATRA 90	87%	73.3%	99.4%	100%
MATRA 130	73.8%	53.9%	96.7%	98.4%
Average	80.4%	63.6%	98.1%	99.2%
Standard deviation	9.0	20.0	1.7	1.7

Table 4
Variance smoothing effects

Method / Database	HMM state variances		HMM fixed variance		HMM smoothed variance	
	Normal	NSS	Normal	NSS	Normal	NSS
MATRA 90	44%	57%	73%	97%	74%	98.6%
MATRA 130	25%	38%	54%	92.7%	54%	96.8%
ENST	25%	60%	61%	94%	64.5%	94.9%
CSELT 70	99%	98.5%	99.5%	99.8%	99.2%	100%
CSELT 130	86.8%	91%	81%	99.8%	83%	99.7%
Average	56.1%	68.7%	73.85%	96.8%	74.8%	98%
Standard deviation	34.5	23.1	23.3	5.5	20.7	2.9

Results with HMMs using the projection or the modified projection

Table 5 summarises the results obtained with the optimal variance smoother for each database using the projection operators. The results obtained with this modified Viterbi scheme are interesting and the gap between the rough results and those obtained with noise compensation is reduced when compared to the standard case (compare Tables 4 and 5).

For each projection measure several weightings have been tested as in the standard Euclidean case: state variances, fixed variance and variance smoothers.

Smoothing effects

In Figures 2 and 3, we report results obtained with the various smoothers tested both for the Euclidean and the projection, using NSS or not. For the exponential lifter, the exponent value has been varied taking values from 0 to 1.

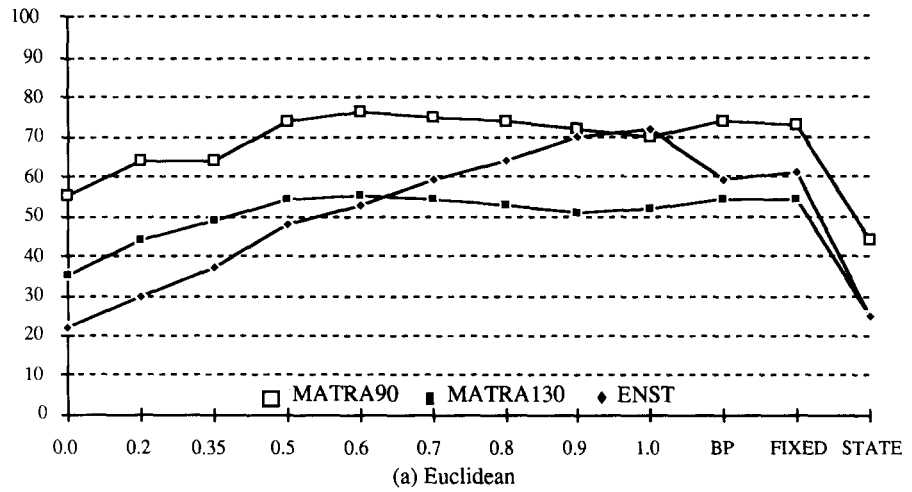
The smoothed or fixed variances clearly outperform the state variances.

The computaton of the SNR is biased due to the fact that the noise conditions are not simulated; anyhow, approximate values of 0 dB on MATRA130 database and −5 dB on the ENST database have been generally observed.

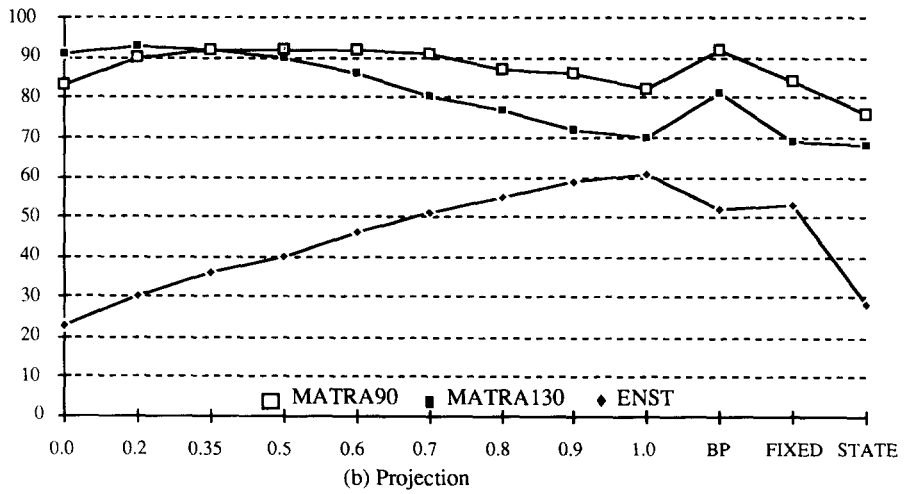
A good compromise can be found with the bandpass lifter (BP); it outperforms the fixed variance in all cases and seems to present a low variability with regard to a fixed exponential lifter: indeed its main advantage resides in the fact that it does not require tuning. Anyhow, on MATRA database, an exponential lifter with exponent ranging between 0.2 and 0.6 is optimal both for the projection and Euclidean cases when NSS is applied. The optimal values are different on the crude results of Figure 2 (shift towards RPS). Comparing Figures 2(a) and 2(b), we observe that the projection significantly improves the results, except for the ENST database which gives degraded performance in this

Table 5
Comparison: HMM and projection/HMM and modified projection

Method / Database	HMM weighted projection	HMM weighted projection modified	HMM weighted projection + NSS	HMM weighted projection mod. + NSS
MATRA 90	89.2%	92%	99%	99.4%
MATRA 130	79%	85.6%	93.7%	98.4%
ENST	46.3%	46%	84.7%	86.6%
CSELT 70	98.7%	99.4%	99.8%	99.8%
CSELT 130	81.2%	87.2%	94.5%	99.4%
Average	79.3%	82.5%	94.5%	96.9%
Standard deviation	20.3	19	6.7	5



(a) Euclidean



(b) Projection

Fig. 2. Euclidean/Projection. No noise compensation. Recognition scores obtained varying the variances with from 0 to 1 the exponential lifter, BP (bandpass), FIXED (fixed variance), STATE (state variances).

case. This could be explained by the low level of the SNR. In Figure 3, the same effect can be observed, though it is considerably reduced due to NSS.

Effect of varying the number of repetitions for training

Figure 4 presents results varying the number of repetitions used for training the system. The variance smoothing effect is here efficient and we get very compact results as compared to the fixed variance.

Linear/nonlinear spectral subtraction (NSS)

Table 6 summarises the results obtained with the nonlinear spectral subtractor. These results also integrate a variance smoother. It can be observed that NSS constantly improves the results over the linear spectral subtraction, this being either with the standard HMM or with the modified Viterbi scheme (HMM + projection).

However the standard HMM with spectral subtraction (linear and NSS) performs slightly better than the modified (HMM with projection), and especially when noise conditions become really difficult (ENST's database). This is particularly

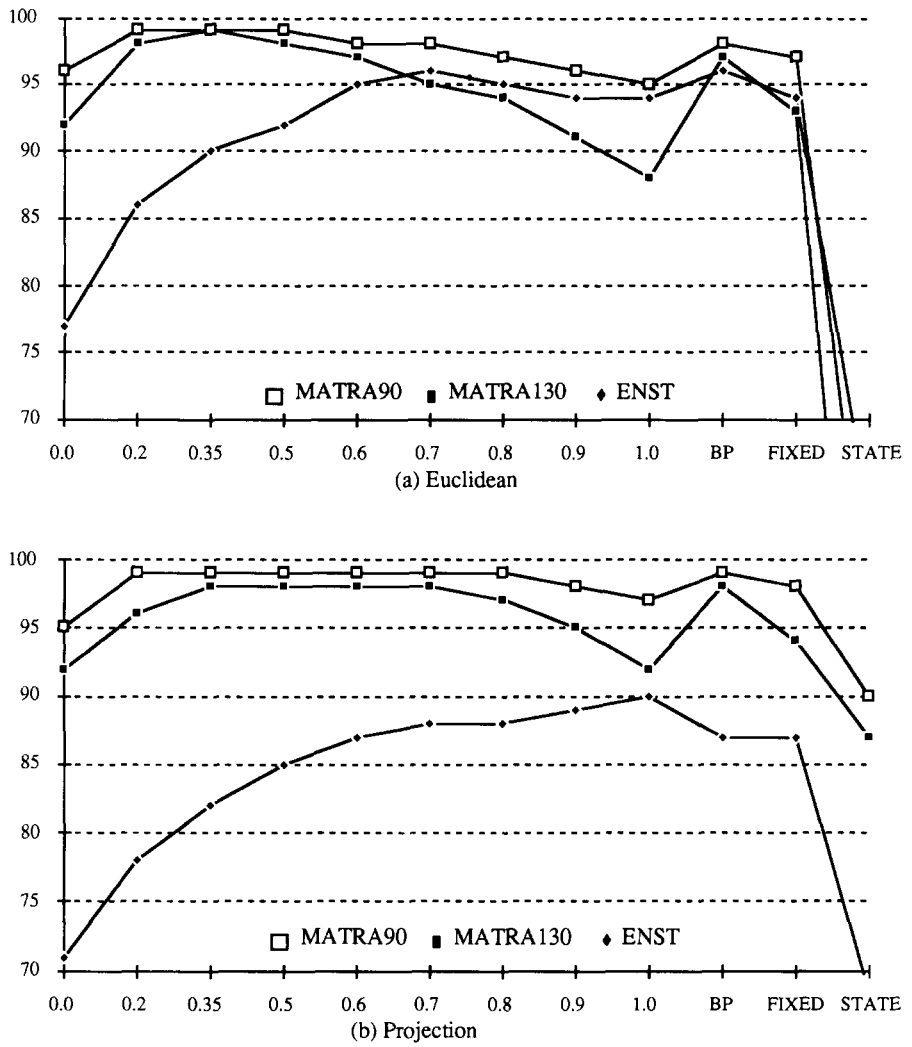


Fig. 3. Euclidean/Projection. NSS. Recognition scores obtained varying the variances with from 0 to 1 the exponential lifter, BP (bandpass), FIXED (fixed variance), STATE (state variances).

materialised by the low variability of the final results with a standard deviation of only 2.8, compared to 5 for the projection.

6. Conclusion

To summarise the results obtained, a number of positive points have been reached and the following results can be put forward.

Nonlinear spectral subtraction. The new noise compensation technique (NSS) can be extended into the HMM framework. NSS outperforms standard spectral subtraction for a DTW system. This difference is even more significant within HMMs. One possible explanation is that the duration model is better represented in a DTW system rather than with HMMs.

Comparisons DTW/HMM. Provided that a good noise compensation is performed (the mismatch between the training and testing conditions is minimised), we find that with the use of NSS, HMMs perform better than DTW-based systems. To be fully objective in this comparison, we should have

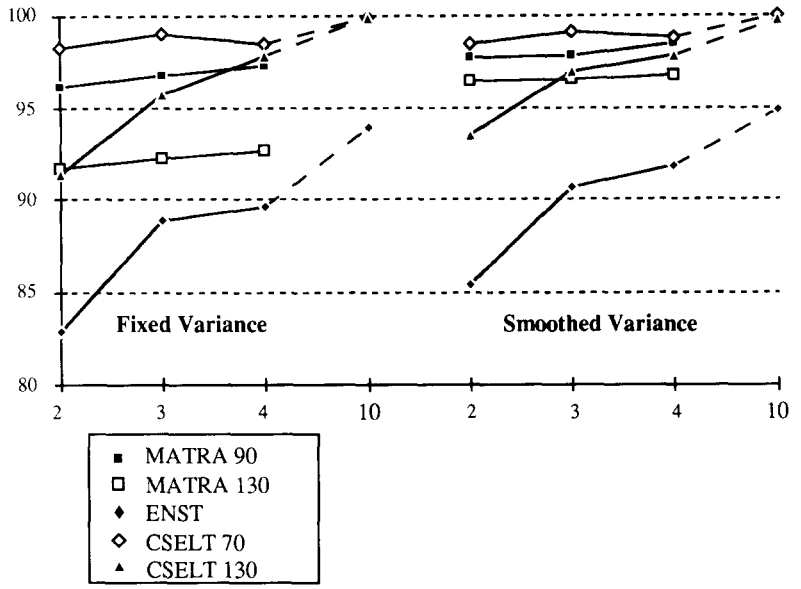


Fig. 4. Effect of varying the number of training utterances for each database considered.

Table 6
Comparisons linear spectral subtraction/NSS for the Euclidean and the modified projection

Method / Database	HMM Smoothed variances (linear spectral subtraction)	HMM Weighted projection modified (linear spectral subtraction)	HMM Smoothed variances (NSS)	HMM Weighted projection modified (NSS)
MATRA 90	96.2%	98.7%	98.5%	99.4%
MATRA 130	93.3%	97.3%	96.8%	98.4%
ENST	90%	81.5%	94.9%	86.6%
CSELT 70	100%	99.7%	100%	99.8%
CSELT 130	98.4%	97%	99.7%	99.4%
Average	95.6%	95%	98%	96.8%
Standard deviation	5.6	6.8	2.8	5

used the DTW system with multiple templates. This was not done to the complexity of the resulting scheme, both from storage and computation points of view.

Robust variances. The fixed variance estimate sometimes performs poorly, probably due to insufficient training material. In this case we have shown that a bootstrap of the variances could be made and, namely, a smoothed variance gives good results.

Euclidean/projection/modified projection. A Viterbi recogniser incorporating the projection measure after adaptation of the density functions of the HMM has been proposed. This system performs better than a standard HMM when no noise preprocessing is performed. The method is significantly improved by the use of NSS. In this case, the modified projection becomes roughly equivalent to the standard Gaussian (Euclidean) case, but with slightly better performances for the latter when using smoothed variances. On the other hand,

Euclidean outperforms projection at very low SNRs, but this effect can be significantly reduced with the use of NSS.

We found also that the modified projection performs better than the standard projection in any conditions, i.e. with and without noise compensation.

Robust HMM system. Finally, the optimal combination of all these optimisations results in efficient HMM systems robust to noise. So by merging the best parametric representation (MFCC), the best noise compensation technique (NSS), the best variance estimate, and a continuous density HMM with optimised topology (16 states, no skip), the following results are obtained both for standard HMMs and for modified HMMs with projection: we observe a rise in performance from 56% to 98% for the HMMs after averaging the results on all noise conditions.

Acknowledgments

The work reported here has been performed within an ESPRIT project P2101 (ARS: Adverse-environments recognition of Speech) sponsored by the CEC. The authors would like to thank the ARS project partners for the availability of their database.

References

C. Baillargeat, I. Lecomte, J. Boudy and P. Lockwood (1990), "Discrete utterance recognition in the car environment", *Proc. ISATA.*

M. Berouti, B. Schwartz and J. Makhoul (1979), "Enhancement of speech corrupted by acoustic noise", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*

S.F. Boll (1979), "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans Acoust. Speech Signal Process.*, Vol. 27, No. 2, April 1979.

J. Boudy, J.F. Dolidet, P. Lockwood and P. Veyrines, Single microphone comparison, Deliverable Task 2200, ARS project, 1990 (unpublished report).

B.A. Carlson and M.A. Clements (1991), "Application of a weighted projection measure for robust hidden Markov model based speech recognition", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*

S.B. Davis and P. Mermelstein (1980), "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-28, pp. 357-366.

G. Faucon R. Le Bouquin, S. Tazi Mezalek, C. Baillargeat, J. Boudy and P. Lockwood (1991), Bilan sur trois méthodes de débruitage de parole, Séminaire Traitement et Représentation du Signal de Parole, Université du Mans (in French).

A. Hermansky and J.C. Junqua (1988), "Optimisation of perceptually-based ASR front-end", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*

B.H. Juang (1991), "Speech recognition in adverse environments", *Comput. Speech Language*, Vol. 5.

B.H. Juang, L.R. Rabiner and J.G. Wilpon (1987), "On the use of band-pass liftering in speech recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-35.

I. Lecomte, M. Lever, J. Boudy and A. Tassy (1989), "Car noise processing for speech input", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*

P. Lockwood and J. Boudy (1992b), "Non-linear spectral subtraction (NSS): A noise compensation technique for speech recognition in car adverse environments", submitted for publication.

P. Lockwood, C. Baillargeat, J.M. Gillot, J. Boudy and G. Faucon (1991), "Noise reduction for speech enhancement in cars: Non-linear Spectral Subtraction/Kalman filtering", *Proc. Eurospeech.*

P. Lockwood, J. Boudy and M. Blanchet (1992a), "Non-linear spectral subtraction (NSS) and Hidden Markov Models for robust speech recognition in car noise environments", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*

D. Mansour and B.H. Juang (1989), "A family of distortion measures based upon projection operation for robust speech recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-37, No. 11.

K.K. Paliwal (1982), "On the performance of the quefrency-weighted cepstral coefficients in vowel recognition", *Speech Communication*, Vol. 1, No. 1, August 1982, pp. 151-154.

D.B. Paul (1980), "The spectral envelope estimation vocoder", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-29, August 1980, pp. 786-794.

D.B. Paul, R.P. Lippmann, Y. Chen, C. Clifford and J. Weinstein (1986), "Robust HMM-based techniques for recognition of speech under stress and in noise", *Proc. SPEECH TECH'86.*

J. Picone (1989), "On modeling duration in context in speech recognition", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*

H.W. Ruehl, S. Dobber, J. Weith, P. Meyer, A. Noll, H.H. Hamer and H. Piotrowski, "Speech recognition in the noisy car environment", *Speech Communication*, Vol. 10, No. 1, February 1991, pp. 11-21.

Y. Tohkura (1987), "A weighted cepstral distance measure for speech recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-35, No. 10.

D. Van Compernolle (1989), "Noise adaptation in a hidden Markov model speech recognition system", Comput. Speech Language, Vol. 3, pp. 151-167.