专 题 汇 编

ICASSP′ 语声识别

北京邮电学院图书馆

AN IMPROVED APPROACH TO THE HIDDEN MARKOV MODEL DECOMPOSITION OF SPEECH AND NOISE

M.J.F. Gales *S.Young*

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, England.

ABSTRACT

This paper addresses the problem of automatic speech recognition in the presence of interfering noise. The new approach described decomposes the contaminated speech signal using a generalisation of standard Hidden Markov Modelling, whilst utilising a compact and effective parametrisation of the speech signal. The technique is compared to some existing noise compensation techniques, using data recorded in noise, and is found to have improved performance compared to existing model decomposition techniques. Performance is comparable to existing noise subtraction techniques, but the technique is applicable to a wider range of noise environments and is not dependent on an accurate endpointing of the speech.

1. INTRODUCTION

The problem of achieving robust speech recognition in noise has attracted a great deal of interest [3, 10, 11][1]. Interfering noise degrades the performance of existing recognition systems, particularly where there is a mismatch in the training and testing environments. Two main approaches to this problem have been studied:-

1. compensation during the data preprocessing stage;
2. compensation during the recognition, or decoding, stage.

The approach described here falls into the second category. The contaminated speech signal is assumed to consist of two independent components, clean speech and contaminating noise, each of which is modelled separately. Some combination of these models is assumed to generate the observed signal.

Model combination is preferable to subtraction based schemes as it does not explicitly rely on the noise power having zero variance. In standard spectral subtraction schemes the case of non zero noise power variance is overcome by setting some spectral subtraction threshold [11]. This threshold is not necessary in a model combination scheme. Furthermore Varga [1, 2] has shown speech and noise signal decomposition to be applicable to scenarios where the noise may be temporally structured and highly time varying e.g. machine gun noise, interfering speech. The technique introduced in this paper uses a compact and effective parametrisation of the signal, Mel-Frequency Cepstrum Coefficients (MFCC) [9]. We combine this parametrisation with a method of model combination, while avoiding some of the assumptions made in previous work [1, 7]. This technique will be referred to as MFCC model combination.

MFCC model combination is evaluated for speech in the presence of car noise using speech recorded in a real car environment. Comparisons in performance are made with *non-linear spectral subtraction* due to Lockwood [8] and the contaminated speech decomposition technique due to Varga [1].

2. THEORY

The observed signal is assumed to consist of two components, noise and clean speech, which are modelled separately. It is assumed that combining the output from these two models will generate the noise contaminated speech signal. The notation used in this paper is that O_t represents the observation in the linear energy frequency domain, O_t^l the log energy frequency domain and O_t^c the energy quefrency domain. A symbol in bold is a vector or matrix, subscripts refer to elements of the vector, hence μ_i refers to the ith element of the vector μ.

Using a Hidden Markov Model (HMM) based recogniser the probability of some observation may be evaluated as

$$\text{Observation Probability} = P(\text{Observation}|\mathcal{M}_1 \otimes \mathcal{M}_2) \quad (1)$$

where $\mathcal{M}_1$ and $\mathcal{M}_2$ are the clean speech model and noise model respectively, trained on MFCC, and $\otimes$ is some combination operator to combine the two models. Recognition is performed using the generalised Viterbi decoding algorithm. This decoder attempts to simultaneously calculate the optimum state sequence in both the speech and the noise models. Probabilities are calculated as

$$P_t(i,j) = \max_{u,v} P_{t-1}(u,v) a1_{u,i} a2_{v,j} b1_i \otimes b2_j(O_t^c) \quad (2)$$

where $P_t(i,j)$ is the probability at time t of being in state i in model 1 and state j in model 2: $a1_{u,i}$ is the transition probability from state u to state i in model 1: $a2_{v,j}$ is the transition probability from state v to state j in model 2 and $b1_i \otimes b2_j(O_t^c)$ is the observation probability. The observation probability takes the general form

$$b1_i \otimes b2_j(O_t^c) = \int P(O1_t^c, O2_t^c|i,j) \quad (3)$$

Where O_t^c is the observed output vector, $O1_t^c$ is the output symbol from model 1, $O2_t^c$ is the output symbol from model 2 and the integral is over the couples $O_t^c = O1_t^c \otimes O2_t^c$. To calculate this probability explicitly over all couples is computationally intractable, so some approximation to the

[1] This work was conducted as part of the Esprit II Project 2101 - Adverse Environment Recognition of Speech

distribution is required. The form of the approximation is dependent on the parametrisation used, in this case MFCC.

The observation sequence in the quefrency domain may be described in terms of the observation sequence in the log energy frequency domain.

$$O_t^c = \mathbf{C}O_t^l \tag{4}$$

where C is the matrix representing the cosine transform, defined as $\mathbf{C}_{ij} = \cos(i(j-0.5)\pi/N)$, where N is the number of energy filterbanks used. It should be noted that the subscripts 1 and 2 are omitted for this section, as the same principle applies to both the noise model and the speech model. From equation 4 it is possible to write

$$O_t^l = \mathbf{C}^{-1}O_t^c \tag{5}$$

The output probability distribution, $b_i(O_t^c)$, is assumed to be gaussian or a mixture of gaussian distributions. Equation 5 describes a linear transformation, therefore the mean, μ^l, and the covariance matrix, Σ^l, can be calculated as

$$\begin{aligned} \mu^l &= \mathbf{C}^{-1}\mu^c && (6)\\ \Sigma^l &= \mathbf{C}^{-1}\Sigma^c(\mathbf{C}^{-1})^T && (7) \end{aligned}$$

Hence the output probability distribution, $b_i(O_t^l)$, is also a single gaussian or mixture of gaussians. For the case of a single gaussian the distribution prior to the log transform is a log normal. The mean may be calculated from

$$\begin{aligned} \mu_i &= \mathcal{E}\left\{e^{x_i}\right\}\\ &= \int\limits_{\mathcal{R}^n} \frac{1}{(\sqrt{2\pi})^n|\Sigma^l|^{\frac{1}{2}}} e^{x_i-\frac{1}{2}(\mathbf{x}-\mu^l)(\Sigma^l)^{-1}(\mathbf{x}-\mu^l)^T}\, d\mathbf{x} \end{aligned}$$

where $\mathcal{R}^n$ is the region of all possible values of x, the observed vector. This equation may be simplified to

$$\mu_i = e^{\mu_i^l+\Sigma_{ii}^l/2} \tag{8}$$

The variance may similarly be calculated.

$$\begin{aligned} &\mathcal{E}\left\{e^{x_i}e^{x_j}\right\}\\ &= \int\limits_{\mathcal{R}^n} \frac{1}{(\sqrt{2\pi})^n|\Sigma^l|^{\frac{1}{2}}} e^{x_i+x_j-\frac{1}{2}(\mathbf{x}-\mu^l)(\Sigma^l)^{-1}(\mathbf{x}-\mu^l)^T}\, d\mathbf{x}\\ &= \mu_i\mu_j e^{\Sigma_{ij}^l} \end{aligned}$$

The variance may then be shown to be

$$\begin{aligned} \Sigma_{ij} &= \mathcal{E}\left\{e^{x_i}e^{x_j}\right\} - \mathcal{E}\left\{e^{x_i}\right\}\mathcal{E}\left\{e^{x_j}\right\}\\ &= \mu_i\mu_j\left[e^{\Sigma_{ij}^l}-1\right] && (9) \end{aligned}$$

From the assumption that the noise power and speech power are independent and additive the combined mean, μ, and covariance, Σ, are given by

$$\begin{aligned} \mu &= \mathrm{SNR}\mu 1 + \mu 2 && (10)\\ \Sigma &= \mathrm{SNR}^2\Sigma 1 + \Sigma 2 && (11) \end{aligned}$$

The SNR term in equations 10 and 11 is required to compensate for both the signal to noise ratio of the observed speech and noise and the mismatch in the training and test energy levels for the speech.

Given the combined mean and variance, and assuming that the combined distribution is approximately lognormal, the combined distribution in the quefrency domain may be obtained by inverting the process previously described. Hence

$$\begin{aligned} \mu_i^l &= \log(\mu) - \frac{1}{2}\log\left[\frac{\Sigma_{ii}}{\mu_i^2}+1\right] && (12)\\ \Sigma_{ij}^l &= \log\left[\frac{\Sigma_{ij}}{\mu_i\mu_j}+1\right] && (13) \end{aligned}$$

Finally converting to the quefrency domain

$$\begin{aligned} \mu^c &= \mathbf{C}\mu^l && (14)\\ \Sigma^c &= \mathbf{C}\Sigma^c\mathbf{C}^T && (15) \end{aligned}$$

The final output probability distribution may be desciibed by

$$b1_i \otimes b2_j(O_t^c) \approx \mathcal{N}(O_t^c, \mu^c, \Sigma^c) \tag{16}$$

where $\mathcal{N}(O_t^c, \mu, \Sigma)$ is a gaussian probability distribution with a mean vector μ and a covariance matrix Σ.

Two problems arise from the use of equation 16. The covariance matrix Σ^c will not in general be a diagonal matrix, this results in an increase in the decoding time. In addition the speech is known to be distorted by the Lombard effect [4, 5]. In [4] various techniques are described for improving HMM robustness to stress distortion, including the Lombard effect. MFCC model combination employs the grand fixed variance, Σ_g^c, in the decoding stage for robustness to stress distortion. Hence the covariance matrix used in the decoder is based on the variance over all the utterances of all the word classes. However the grand fixed variance may not be used in the calculation of the combined means and variances. The grand variance is an over-estimate of the actual variance, hence its use in equation 8 would result in an over-estimate of the mean in the linear domain. The actual output probability distribution used is

$$b1_i \otimes b2_j(O_t^c) \approx \mathcal{N}(O_t^c, \mu^c, \Sigma_g^c) \tag{17}$$

Using equation 17 also ensures that the covariance matrix is diagonal.

Equation 17 may be compared with the output probability proposed by Varga [1]

$$\begin{aligned} b1_i \otimes b2_j(O_t^l) \approx\ & \mathcal{C}(O_t^l, \mu 1^l, \Sigma 1^l)\mathcal{N}(O_t^l, \mu 2^l, \Sigma 2^l)\\ &+ \mathcal{N}(O_t^l, \mu 1^l, \Sigma 1^l)\mathcal{C}(O_t^l, \mu 2^l, \Sigma 2^l) && (18) \end{aligned}$$

and the standard output probability distribution using a single model

$$b1_i(O_t^c) \approx \mathcal{N}(O_t^c, \mu 1^c, \Sigma 1^c) \tag{19}$$

where $\mathcal{C}(O_t, \mu, \Sigma)$ is the cumulative probability function with mean vector μ and covariance matrix Σ. In both equation 18 and equation 19 the covariance matrices are assumed to be diagonal.

The use of MFCC model combination also reduces the problem of models learning low energy events, which are visible in the clean training environment, but are not visible in the test environment [11]. This aspect has been compensated for in spectral subtraction schemes by the addition of artificial noise in both the training and test environments. However the artificial noise added in these situations is subjectively chosen a few dB above the noise floor, possibly resulting in unnecessary masking of some useful speech features. By combining the models in the linear-energy domain low energy speech events are automatically masked in the presence of high energy noise.

3. COMPUTATIONAL COMPLEXITY

A critical aspect in the use of any noise compensation scheme is the computational overhead associated with its implementation. This overhead may result from additional complexity in both the preprocessing and the decoding stages of the recogniser.

The use of standard spectral subtraction schemes [6] result in additional complexity at the preprocessing level, due to the estimation of the noise power spectrum and the subtraction process. There is no additional overhead associated with the decoding of the speech.

For the noise compensation scheme of Varga, overheads are associated with both the pre- and the post-processing stages. Firstly a noise model must be correctly trained and updated as the noise environment changes. It would therefore be desirable to use a global noise model, trained over a variety of noise conditions. In addition there is an overhead associated with the decoder. Comparing equation 18 and equation 19 an increase by a factor of four in complexity may be seen. There is also the fact that the technique is based on a log-energy filterbank frontend. This representation has been shcwn to be not as informationally intensive as cepstral coefficients [9]. Hence the number of parameters required is generally greater than for MFCC parametrisation.

MFCC model combination has the same requirement that a noise model be calculated, as in the Varga compensation scheme. Furthermore the model combining process is computationally expensive. However by using a global noise model for the short term variations in the noise, the rate of update for combining models and training new noise models is greatly reduced. For a single state noise model there is no decoder overhead, compare equation 17 and equation 19, however for more complex noise models a decoder overhead is incurred. If an n state noise model, $\mathcal{M}_n$, is combined with an m state speech model, $\mathcal{M}_m$, the combined model, $\mathcal{M}_m \otimes \mathcal{M}_n$, is an $m * n$ state model. The factor of n increase in the number of states results in a factor of n increase in the decoder overhead. This additional overhead is associated with any model combining process, including the technique of Varga.

4. RESULTS

4.1. Databases

The databases used for the evaluation of the various techniques were collected for ESPRIT II project No. 2101, Adverse Environment Recognition of Speech.

The main database used was the ENST-1 ARS database. This database consists of a 43 word vocabulary, based on a possible set of in-car telephone commands, 10 training utterances recorded in a stationary car and 15 word utterances recorded during normal motorway driving. The speaker, yg, was male, using his native language, French. The signal to noise ratio (SNR) for the database was 0.52dB.

The second database used, the ENST-2 ARS database, consists of a 43 word vocabulary recorded under the same conditions as the first database. A different speaker, em, was used, again male, speaking his native tongue, French. The SNR for the second database was −3.53dB.

4.2. HMM Recogniser

The baseline recogniser was a 10 state word based HMM, with eight emitting states and single gaussian output probability distributions. The speech was parametrised using the first 10 MFCC, ignoring the zeroth coefficient. The two standard spectral subtraction schemes, non-linear spectral subtraction (NSS), due to Lockwood [8], and linear spectral subtraction (LSS) based on the enhancement scheme of Boll [10], both use a local estimate of the noise mean. In addition NSS uses a local noise maximum in the noise compensation scheme.

For MFCC model combination the speech models, again 10 state word models with eight emitting states, were trained on eleven cepstral coefficients, including the zeroth coefficient. The noise model used, a single emitting state model, was trained on all the available noise data. After combining, the decoding was based on the same 10 cepstral coefficients as the baseline decoders using the expression for the output probability distributions shown in equation 17.

The approach described by Varga [1] was also implemented using a 27 log energy filterbank, based on a mel scale, for the parametrisation of the speech. 10 state word models were used for the speech and a single emitting state model was used for the noise. Decoding was performed using the complete 27 parameter vector.

4.3. Recognition Results

Technique	*Recognition rate*
No Compensation	49.15
Standard LSS	84.34
Matra NSS	92.40

Table 1: Baseline Recognition Results

The baseline recognition rates for the standard spectral subtraction techniques are shown in table 1.

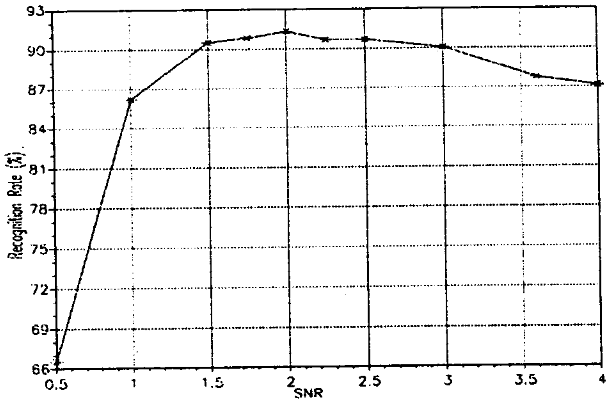

Figure 1: Recognition rates vs assumed SNR for ENST-1

Figure 1 shows the recognition rates against assumed SNR for MFCC model combination. The best recognition rate is 91.32% at an assumed SNR of 2.0. This result is comparable with the recognition rate achieved by non-linear spectral subtraction, 92.40%.

The mismatch between the optimal decoder SNR and the actual SNR of the test data is partly due to the difference in the energy levels between the training and the test speech. This mismatch must be compensated for, as the models are combined in the linear energy domain, which is sensitive to alterations in absolute energy levels. Examining the total speech gain between training and testing the SNR was found to be 3.6 (5.56dB), this corresponds to a recognition rate of 87.75%. The use of a lower than theoretically predicted value of SNR may be accounted for by the effect of stress on the cepstral coefficients. In [5] the cepstral norm of lombard speech vowels is found to decrease by 15% to 30% and energy levels at various frequency bands to decrease. From figure 1 we see that the performance is not critically dependent on the value of SNR chosen, since recognition rates are over 90% for the range $1.5 \leq \text{SNR} \leq 3.0$.

SNR	*Recognition rate*
2.0	69.61
3.6	54.57

Table 2: Varga Model Combination Recognition Rates

Recognition rates for the Varga decoder are shown in table 2. As the Varga decoder is dependent on the absolute values of the speech energy it was necessary to use an assumed SNR, the values used being the theoretical value and the 'optimal' model combining value previously found. Again the theoretical value is found not to produce the best results. The results show improvement over the standard baseline decoder, but the performance is not comparable to the standard noise compensation schemes.

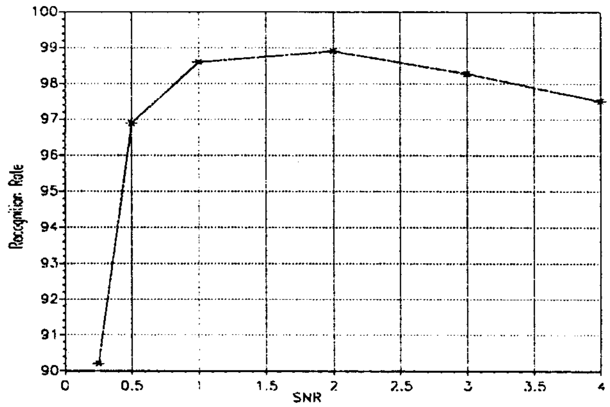

Figure 2: Recognition rates vs assumed SNR for ENST-2

The two most successful techniques on the ENST-1 ARS database, NSS and MFCC model combination, were tested using the ENST-2 ARS database. The models were trained as previously described. Figure 2 shows recognition rates against assumed SNR for the ENST-2 database. The comparative results are shown in table 3.

Technique	*Recognition rate*
No Compensation	53.65
NSS	94.87
Best MFCC comb.	98.91

Table 3: Recognition Rates for ENST-2

A theoretical value for the assumed SNR is calculated to be -0.28dB (0.9367), again this does not correspond to the 'optimal' combination value of 3.0dB (2.0). However from the high recognition rates, the speaker stress may be assumed to be lower than that of the ENST-1 database. Hence the theoretical value for the SNR results in recognition performance not significantly lower than that of the 'optimal' value. The performance is superior to that of NSS.

5. CONCLUSIONS

In this paper we have examined a new technique, MFCC model combination, for the decomposition of speech in noise. The technique has been shown to produce greater improvements in recognition rate than that of Varga [1] and comparable or better results than non-linear spectral subtraction. Furthermore the proposed technique uses a global noise model, which cannot be employed in non-linear spectral subtraction. The use of a global noise estimate reduces the need for accurate end-pointing during recognition. No computational overhead is associated with the use of MFCC model combination at the decoder stage, though a noise model is required to be trained and updated.

Presently the technique has only been applied in a scenario where a simple noise model is appropriate. Where the noise has strong temporal structure the technique should show significantly better performance than standard spectral subtraction techniques. This aspect will be investigated in future work.

REFERENCES

[1] Varga A.P. and Moore R.K. Hidden markov model decomposition of speech and noise. In *Proceedings ICASSP*, pages 845–848, 1990.

[2] Varga A.P. and Moore R.K. Simultaneous recognition of concurrent speech signals using hidden markov model decomposition. In *Proceedings Eurospeech*, 1991.

[3] Mansour D and Juang BH. The short-time modified coherence representation and noisy speech recognition. In *IEEE Trans. Acoust., Speech Signal Processing*, volume 37, pages 795–804, 1989.

[4] Paul DB. A speaker-stress resistant hmm isolated word recogniser. In *Proceedings ICASSP*, pages 713–716, 1987.

[5] Junqua J and Anglade Y. Acoustic and perceptual studies of lombard speech: Application to isolated-word automatic speech recognition. In *Proceedings ICASSP*, pages 841–844, 1990.

[6] Lim JS and Oppenheim AV. Enhancement and bandwidth compression of noisy speech. In *Proceedings IEEE*, volume 67, pages 1586–1604, 1979.

[7] Wang M and Young S. Speech recognition using hidden markov model decomposition and a general background speech model. In *Proceedings ICASSP*, 1991.

[8] Lockwood P. and Boudy J. Experiments with a non linear spectral subtractor (nss), hidden markov models and the projection, for robust speech recognition in cars. In *Proceedings Eurospeech*, 1991.

[9] Davis S.B. and Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *IEEE Transactions ASSP*, volume 28, pages 357–366, 1980.

[10] Boll SF. Suppression of acoustic noise in speech using spectral subtraction. In *IEEE Transactions ASSP*, volume 27, pages 113–120, 1979.

[11] van Compernolle D. Noise adaptation in the hidden markov model speech recognition system. In *Computer Speech and Language*, volume 3, pages 151–168, 1987.

SPEECH RECOGNITION IN NOISE USING A PROJECTION-BASED LIKELIHOOD MEASURE FOR MIXTURE DENSITY HMM'S

Beth A. Carlson and Mark A. Clements

School of Electrical Engineering
Georgia Institute of Technology
Atlanta, GA 30332

ABSTRACT

In this study, a cepstral likelihood measure based upon the projection operation is incorporated into a mixture density HMM scheme to improve recognition in the presence of additive noise. The case is addressed where the models are determined only under noise-free conditions. A background discussion and derivation of the measure is provided. Recognition experiments are presented showing the usefulness of the proposed measure over the standard Gaussian measure (weighted Euclidean distance) for speaker independent, isolated word recognition in noise. It was found that the proposed *mixture weighted projection measure* significantly improved performance in several noise types, including white, jittering white, and colored noise. As an example, at an SNR of 10 dB white noise, recognition improved from only 38.4% correct using the Gaussian measure to 83.6% using the developed measure.

1. INTRODUCTION

A speech recognizer designed to perform well under noise-free conditions usually will show marked degradation in performance when background noise is present. For the most part, methods presented in the literature to improve recognition in noise tend to focus on ways to effectively "remove" the noise from the degraded speech. This can be in the form of enhancement techniques such as spectral subtraction or microphone arrays which try to filter out the noise or in the form of modeling techniques which try to find models for the noise-free speech and the noise signal, such as in HMM's [1]. However, such methods have shown limited performance improvements and can only be implemented under conditions where an estimate of the noise statistics is known. The approach taken in this paper requires no estimate of the noise statistics and does not require that they be constant. The focus is on developing likelihood measures which are robust to the effects of noise rather than trying to remove the noise from the degraded speech. A similar approach has been taken more recently in the literature for finding both distance measures, such as spectral slope [2] and weighted Itakura-Saito distance [3], and robust feature representations, such as the IMELDA system [4], which are to a certain extent less affected by noise.

In our previous work [5], it was shown how a likelihood measure based upon the projection operation could be incorporated into the Viterbi algorithm for the case of continuous density HMM's. The measure was developed for both the cepstral and mel-cepstral representations for speech. The proposed measure was found to greatly improve speaker dependent, isolated word recognition performance in the presence of white noise. Hence, in our present work, we extend the use of this projection-based likelihood score for use in mixture density HMM's for speaker independent recognition. In addition, the performance of the measure is evaluated for other noise types, including white, jittering white, and broadband colored noise to show its versatility in varying backgrounds.

This paper is organized as follows. First, a background discussion of the formulation of the original measure is presented, along with the derivation of the measure for mixture density HMM's. Following this, recognition results are presented using the developed mixture weighted projection measure for recognition in varying noise types. Last, a summary of the results is given.

2. DERIVATION

2.1. BACKGROUND

The original formulation of the projection measure as a distance between two cepstral vectors was introduced by Mansour and Juang [6]. The measure was based upon both theoretical and empirical observations of how white Gaussian noise affects the cepstral coefficients. It was found that the norm of these vectors tended to decrease as more white noise was added and that the angle between the vectors was less affected by the noise. The idea in our previous work was to formulate a likelihood measure for HMM's which would compensate for the effects of the noise in a similar fashion

0-7803-0532-9/92 $3.00 © 1992 IEEE

to the measure of Mansour and Juang.

The modification was made within the likelihood scores of the Viterbi algorithm. Each word model is defined by the parameter set $\lambda = (\pi, A, B)$, where π is the initial state vector, A is the transition matrix, and B is the collection of b_i, which defines the observation probability density function of the i^{th} state in the model. These state distributions are assumed to be Gaussian functions of the following form:

$$\begin{aligned} b_i(\mathbf{c}_t) &= \mathcal{N}(\mathbf{c}_t, \mu_i, \mathbf{C}_i) \\ &= (2\pi)^{-\frac{N}{2}} |\mathbf{C}_i|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{c}_t - \mu_i)^T \mathbf{C}_i^{-1} (\mathbf{c}_t - \mu_i)\right) \end{aligned}$$

where $\mathbf{c}_t$ is the test cepstral vector, μ_i is the mean cepstral vector of the i^{th} state in the model and $\mathbf{C}_i$ is its corresponding covariance matrix. The standard negative log likelihood measure used can be defined as a *Gaussian* or *weighted Euclidean distance*:

$$d_{WE}(\mathbf{c}_t, b_i) = (\mathbf{c}_t - \mu_i)^T \mathbf{C}_i^{-1} (\mathbf{c}_t - \mu_i) + \log |\mathbf{C}_i| \ , \qquad (1)$$

The modified likelihood score, known as the *weighted projection measure*, was developed as follows:

$$\begin{aligned} d_{wproj}(\mathbf{c}_t, b_i) &= (\mathbf{c}_t - \lambda\mu_i)^T \mathbf{C}_i^{-1} (\mathbf{c}_t - \lambda\mu_i) + \log |\mathbf{C}_i| \\ &= |\mathbf{c}_t|^2 (1 - \cos^2 \beta) + \log |\mathbf{C}_i| \ , \qquad (2) \end{aligned}$$

where $\cos \beta$ is the angle between the vectors $\mathbf{c}_t$ and μ_i weighted in the space of $\mathbf{C}_i^{-1}$ and λ is the scale factor defined as,

$$\lambda = \frac{\mathbf{c}_t^T \mathbf{C}_i^{-1} \mu_i}{\mu_i^T \mathbf{C}_i^{-1} \mu_i} \ .$$

This scale factor was incorporated to compensate for the norm reduction and is equivalent to the projection of the noisy test vector, $\mathbf{c}_t$, onto the noise-free reference, $\mathbf{c}_r$ in the weighted space of $\mathbf{C}_i^{-1}$. A further discussion of the theoretical interpretation of the measure and its use for other parameter sets including the mel-cepstral parameters used in this paper can be found in [5]. When used in a HMM system trained with noise-free speech, the weighted projection measure defined above was found to greatly improve recognition performance over the standard Gaussian likelihood measure defined in Equation 1 for speaker dependent, isolated word recognition.

2.2. EQUATION FORMULATION

The concept of the weighted projection measure defined above for uni-modal Gaussian state distributions can be expanded in a straightforward manner to the case where the densities are mixtures of Gaussian functions. The form of the mixture densities used in this study are based upon a partitioning of the parameter space, where each Gaussian function in the mixture represents a partition or cluster in the parameter space. Using this mixture partitioning, the most likely mixture in each state is chosen for each observation, with the likelihood of the i^{th} state defined as

$$b_i(\mathbf{c}_t) = \frac{1}{M} \max_{1 \le k \le M} \mathcal{N}(\mathbf{c}_t, \mu_{ik}, \mathbf{C}_{ik}) \ ,$$

where μ_{ik} and $\mathbf{C}_{ik}$ are the mean and covariance of the k^{th} mixture in the i^{th} state, N is the order of the observation vectors, and M is the number of mixtures (assumed the same for all states). When using the Viterbi algorithm to determine the probability that a given observation sequence was produced by the model, the log likelihood of the above probability is needed. For the partitioned mixture above, the total log likelihood will be dominated by the minimum negative log probability of all the Gaussian functions in the mixture (omitting the $1/M$ term):

$$\begin{aligned} \log b_i(\mathbf{c}_t) &= \log \left[\max_{1 \le k \le M} \mathcal{N}(\mathbf{c}_t, \mu_{ik}, \mathbf{C}_{ik}) \right] \\ &= -\frac{1}{2} \min_k \left[(\mathbf{c}_t - \mu_{ik})^T \mathbf{C}_{ik}^{-1} (\mathbf{c}_t - \mu_{ik}) + \log |\mathbf{C}_{ik}| \right] . \end{aligned}$$

The total negative log likelihood can then be expressed in the form of a *mixture weighted Euclidean distance* between an observation and a mixture density:

$$d_{MWE}(\mathbf{c}_t, b_i) = \min_{1 \le k \le M} [d_{WE}(\mathbf{c}_t, b_{ik})] \ . \qquad (3)$$

The distance $d_{WE}(\cdot)$ is the weighted Euclidean distance for a single Gaussian defined in Equation 1.

To develop the projection measure for the mixture densities, the effects of noise on each Gaussian in the mixture can be considered independent, so the same partitioning as used above can still be assumed. A similar assumption on the effects of noise was used in the noise removal system of Ephraim [1] with good results. As with the uni-modal Gaussian HMM's, a scale factor can be incorporated into the distribution function which compensates for the norm reduction caused with added noise. This results in a modified likelihood expressed as:

$$\tilde{b}_i(\mathbf{c}_t) = \frac{1}{M} \max_{1 \le k \le M} \mathcal{N}(\mathbf{c}_t, \lambda_k \mu_{ik}, \mathbf{C}_{ik})$$

where λ_k is the scale factor for the k^{th} mixture of the i^{th} state in the model. The covariance matrix is not scaled here because in practice it can distort the parameter space too much and result in poor estimates of the variance. Weighting the covariance by λ^2 and then solving for the optimal λ values resulted in very poor recognition performance in preliminary testing in this study and therefore was not pursued further.

Again, the log likelihood of the above probability is needed (omitting the $1/M$ term):

$$\begin{aligned} \log \tilde{b}_i(\mathbf{c}_t) &= \log \left[\max_{1 \le k \le M} \mathcal{N}(\mathbf{c}_t, \lambda_k \mu_{ik}, \mathbf{C}_{ik}) \right] \\ &= -\frac{1}{2} \min_k \left[(\mathbf{c}_t - \lambda_k \mu_{ik})^T \mathbf{C}_{ik}^{-1} (\mathbf{c}_t - \lambda_k \mu_{ik}) + \log |\mathbf{C}_{ik}| \right] . \end{aligned}$$

From the orthogonality principle, the optimal value for λ_k can be found on a frame-by-frame basis by assuming that each mixture is independent of the others:

$$\lambda_k = \frac{\mathbf{c}_t^T \mathbf{C}_{ik}^{-1} \mu_{ik}}{\mu_{ik}^T \mathbf{C}_{ik}^{-1} \mu_{ik}} \ .$$

With these values for the λ_k's, the negative log likelihood can be expressed in the following equation and will be referred to as the *mixture weighted projection measure*:

$$d_{Mwproj}(\mathbf{c}_t, b_i) = \min_{1 \leq k \leq M} [d_{wproj}(\mathbf{c}_t, b_{ik})] \ . \qquad (4)$$

The distance $d_{wproj}(\cdot)$ is the previously used weighted projection measure for a single Gaussian function, defined in Equation 2. For computational purposes, this measure can be expressed in terms of the angle between the observation vector and the state mean in the weighted Euclidean space of $\mathbf{C}_{ik}^{-1}$ (as defined in Equation 2).

Thus, it has been shown how the weighted projection measure can be integrated into mixture density HMM's by simplying altering the likelihood scores within an existing HMM recognition system. The results reported in this paper evaluate the effectiveness of the developed measure for speaker independent recognition in noise.

2.3. IMPLEMENTATION ISSUES

Several issues involved in the actual implementation of the mixture weighted projection measure should be pointed out.

1. The HMM model training is performed using only the standard mixture Gaussian measure (as defined in Equation 3).
2. Either the mixture weighted Gaussian measure or the mixture weighted projection measure is used for computing the likelihood scores in the Viterbi algorithm.
3. When applied to an observation vector consisting of both static and differential parameters, the weighted projection measure is applied to each feature set separately. That is,

$$\begin{aligned} d_{wproj}(\mathbf{o}_t, b_i) \quad = \quad & d_{wproj}(\mathbf{c}_t, b_{i,static}) + \\ & d_{wproj}(\delta_t, b_{i,delta}) \ , \end{aligned}$$

where $\mathbf{c}_t$ and δ_t are the set of static and delta cepstral or mel-cepstral parameters, respectively, with corresponding state densities, $b_{i,static}$ and $b_{i,delta}$.

3. RECOGNITION EXPERIMENTS

Several recognition experiments were conducted to evaluate the effectiveness of the weighted projection measure with mixture density HMM's for speaker independent recognition in noise. A subset of the Texas Instruments Isolated Digits Database was used consisting of the digits "one"–"nine" and the word "oh" with 65 male speakers uttering each word twice. The HMM word models were trained using noise-free speech from 40 male speakers for a total of 80 training tokens per word. The test set consisted of the remaining 25 speakers for a total of 50 tokens per word. Originally sampled at 20 kHz, the speech was filtered and downsampled to 10 kHz. The speech analysis was performed on 30 msec frames of speech every 15 msec using a Hamming window. Extracted parameters included 16 mel-cepstral and 16 delta mel-cepstral parameters computed from a 512-point DFT power spectrum [5]. Noisy speech was simulated by adding artificially generated noise to the speech waveforms at various global signal-to-noise ratios.

The HMM models were left-to-right consisting of 5 states with 5 mixtures per state. Each mixture Gaussian was defined by a mean vector of 16 mel-cepstral and 16 delta mel-cepstral parameters with a corresponding diagonal covariance matrix. Further details of this system can be found in [7].

The following set of experiments evaluates the recognition performance in various noise types, including white, jittering white, and broadband colored noise. The performance of three methods are evaluated: (1) the standard procedure using the mixture Gaussian likelihood measure (as defined in Equation 3) with noise-free training, (2) the modified procedure using the mixture weighted projection measure (as defined in Equation 4) with noise-free training, and (3) the limiting performance of the system using the mixture Gaussian measure with word models derived from noisy speech.

3.1. RECOGNITION IN WHITE NOISE

As can be seen in Figure 1, the mixture weighted projection measure significantly improved the recognition performance over all SNR's of added white noise. An average of 10 dB in improvement was achieved, with error rates reduced by 30 to 80 percent. For example, at 15 dB SNR, recognition improved from only 61% correct using the standard measure to 94.7% correct using the developed measure. In addition, the performance of the mixture weighted projection measure was found to be comparable to that of training and testing in noise for SNR's of 15 dB and above.

3.2. RECOGNITION IN JITTERING NOISE

In this next set of experiments, the speech waveforms were degraded by additive white noise with jittering levels. This noise was generated by randomly varying the levels of white noise from a global SNR of 0 to 20 dB on short-time intervals. This type of noise addresses the problem of non-stationarity in a noisy environment which the previous noise signal did not. The recognition accuracy was 73.8% for the mixture weighted projection measure as compared to only 30% using the mixture weighted Euclidean measure. This represents a cut in the error rate by one-half. However, the approach did not perform as well as training the HMM word models using speech degraded with jittering white noise, which achieved 95.5% accuracy.

3.3. RECOGNITION IN COLORED NOISE

The last set of trials were conducted for speech plus added broadband colored Gaussian noise. The additive noise signal was generated from a 2-pole AR process with filter coefficients $a_1 = -0.45$ and $a_2 = 0.55$ (with a spectral peak at 1 kHz). As can be seen in Figure 2, the developed measure significantly outperformed the standard measure. For an example, recognition accuracy at 15 dB SNR improved from 77.5% correct using the standard measure to 94.2% correct using the mixture weighted projection likelihood measure. Also, for SNR's of 15 dB and above, the performance of the mixture projection measure with noise-free training compared favorably with the limiting case of training and testing in the same noisy environment.

4. SUMMARY

This paper extended the use of the weighted projection measure developed in [5] to the case where the HMM state densities are mixtures of Gaussian functions. The performance of the developed likelihood measure was evaluated for speaker independent digit recognition in the presence of three noise types: white, jittering white, and colored noise. In all cases, the mixture weighted projection measure was found to significantly outperform the standard weighted Euclidean likelihood distance in recognition trials.

The proposed method offers several advantages over other approaches. First, unlike many other methods, the approach here does not require an explicit estimate of the noise and is robust over a wide range of SNR's and noise types. Second, the approach offers a relatively simple modification to existing HMM-based systems which shows great improvements for recognition in noise. Last, the effectiveness of the enhancement technique should translate to other recognition tasks, such as keyword spotting and continuous speech as well.

5. REFERENCES

[1] Y. Ephraim, "A minimum mean square error approach for speech enhancement," *Proc. Int. Conf. Acoust.,Speech,Signal Processing*, pp. 829–832, 1990.

[2] B. Hanson and H. Wakita, "Spectral slope distance measures with linear prediction analysis for word recognition in noise," *IEEE Trans. Acoust.,Speech,Signal Processing*, vol. ASSP-35, pp. 968–973, July 1987.

[3] P. Chu and D. Messerschmitt, "A frequency weighted Itakura-Saito spectral distance measure," *IEEE Trans. Acoust.,Speech,Signal Processing*, vol. ASSP-30 pp. 545–560, Aug. 1982.

[4] M. Hunt, S. Richardson, D. Bateman, and A. Piau, "An investigation of PLP and IMELDA acoustic representations and of their potential for combination," *Proc Int. Conf. Acoust.,Speech,Signal Processing*, pp. 881–884, 1991.

[5] B. Carlson and M. Clements, "Application of a weighted projection measure for hidden Markov based speech recognition in noise," *Proc. Int. Conf Acoust.,Speech,Signal Processing*, pp. 921–924, 1991.

[6] D. Mansour and B. Juang, "A family of distortion measure based upon projection operation for robust speech recognition," *IEEE Trans. Acoust.,Speech,Signal Processing*, vol. ASSP-37, pp. 1659–1671, Nov. 1989.

[7] B. Carlson, *A Projection-Based Measure for Automatic Speech Recognition in Noise.* Ph.D. thesis, Georgia Institute of Technology, Nov. 1991.

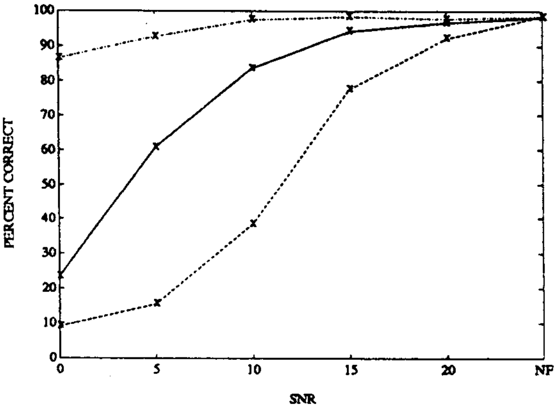

Figure 1: Recognition results for speech in white noise using the Gaussian measure (- - -), the mixture weighted projection measure (—), and training in noise (-·-·-).

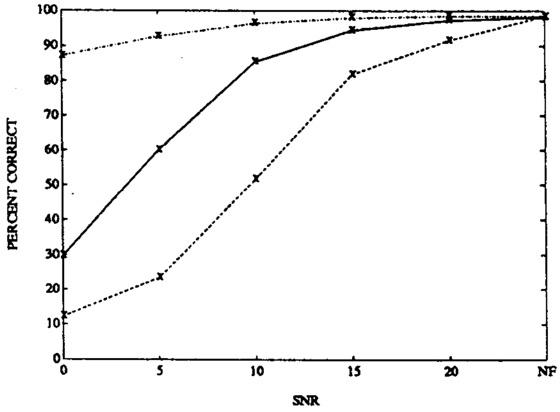

Figure 2: Recognition results for speech in colored noise using the Gaussian measure (- - -), the mixture weighted projection measure (—), and training in noise (-·-·-).

SPECTRAL CONTRAST NORMALIZATION AND OTHER TECHNIQUES FOR SPEECH RECOGNITION IN NOISE

D.C. Bateman, D.K. Bye and M.J. Hunt

Marconi Speech & Information Systems
Airspeed Road, The Airport
Portsmouth, PO3 5RE
England

ABSTRACT

Three methods of improving speech recognition in noise are considered: energy thresholding, a noise-robust spectral representation called IMELDA, and a set of noise-robust spectral distortion measures. The spectral distortion measures can be seen as normalizing the contrast in the spectrum, a property which can be transferred to the representation itself, making it computationally more efficient. In speaker-independent alphabet recognition tests in added steady white noise at various levels, IMELDA is shown to outperform a weighted cepstrum representation and be computationally more efficient. With this material and with digits recorded in trucks at a wide range of noise levels, performance is found to depend strongly on the threshold level. Contrast normalization is found to help, but only when the energy threshold is far from its optimum level.

1. OVERVIEW OF TECHNIQUES

Techniques that have proved helpful for speech recognition in noise with systems using DTW or HMM with continuous-valued parameters include the following:

i) Spectral Thresholding and Subtraction
For systems with filter-bank front-ends, thresholding [1] sets a lower limit on the energy in each channel for both the reference speech and the speech to be recognized (the "test speech"). A threshold is applied explicitly or implicitly in all systems using a log-energy representation, but ideally it should be close to the typical noise level. This level therefore needs to be reasonably constant and to be known approximately. Spectral subtraction [2] — subtraction of an estimate of the current noise spectrum from the noisy speech spectrum — is more dependent than thresholding on the noise being constant and accurately known.

ii) Noise-Robust Acoustic Representations
A representation known as IMELDA [3,4], a linear transformation of the output of a filter-bank, has been derived statistically from a combination of undegraded data and data degraded with steady white noise and with filtering causing spectral tilt changes. It has been shown to improve recognition performance in both the undegraded and degraded conditions. There were reasons to believe it capable of dealing with non-steady and non-white noise of unknown level, but few direct tests of this assertion had been published.

iii) Noise-Robust Spectral Distortion Measures
Certain spectral distortion measures [5] — loosely, distance measures — have been shown to be less affected by the addition of steady white noise to the speech than the standard Euclidean distance. A particular advantage of these measures is that the noise level need not be known.

This paper describes experiments in combining spectral thresholding with IMELDA and with a technique in the spirit of the noise-robust distance measures. The frontier between acoustic representations and distance measures is not sharp; and to provide a computationally efficient combination of IMELDA with the properties of a noise-robust distance measure, we have transferred those properties to the representation. The experiments include tests with varying and non-white noise.

2. SPECTRAL CONTRAST NORMALIZATION

Mansour and Juang [5] pointed out that when white noise is added to speech the relative values of (LPC) cepstrum coefficients are less affected than their overall magnitude as measured by their sum of squares or *norm.* They also pointed out that frames with high norms are less affected than those with low norms.

The cosine transform, being orthonormal, preserves the norm of a frame of spectral data. The norm in the cepstrum domain is therefore equal to the norm in the log spectrum domain. The exclusion of C_0 from the cepstral norm is equivalent to subtracting the mean from the log spectrum. The norm is then equivalent to the spectral variance across a frame, which we call *spectral contrast.* White noise will tend to reduce the dynamic range across the frame and hence the spectral contrast, but will not affect the locations of spectral peaks, which

0-7803-0532-9/92 $3.00 © 1992 IEEE

tend to be reflected in the relative values of cepstral coefficients. Since frames with low contrast correspond to sounds such as voiceless fricatives with little spectral structure, it is not surprising that they are particularly badly affected by noise.

Mansour and Juang compared several distance measures in speaker-dependent and independent tests. In one such measure (which they called d_2) the norm of the model state centroid was adjusted to minimize its Euclidean distance from the test frame to which it was being matched. In another (d_3) each test frame and state centroid was scaled to have a norm of unity; and in a third (d_5) the squared distance d_3 was scaled by the norm of the test frame, so giving greater weight to the more reliable high-norm frames. All three outperformed the standard Euclidean distance on speech degraded with steady white noise, with d_5 being the best.

Subsequently, Carlson and Clements [6] used d_2 in speaker-dependent experiments with a mel-cepstrum derived from a DFT and added dynamic parameters to the static set.

3. CONTRAST NORMALIZED FRONT-END

IMELDA is computationally efficient because the matrix multiplication needed to implement it is applied in the front-end once per frame rather than in the comparison process between the test frame and the model states, which would entail many matrix multiplications per test frame. Since the contrast-normalizing distance measure d_2 used by Carlson and Clements requires knowledge of both the test frames and the state centroids, the contrast normalization and the cepstral weighting or IMELDA transformation must be carried out as part of the state-to-frame comparison process. The distance measure d_5 is more attractive, since the contrast normalization and the IMELDA matrix multiplication can be carried out in the front-end, adding only a scaler multiplication to the frame-to-frame comparison. However, with our specialized hardware designed for unweighted Euclidean distances, this slight complication would have serious consequences.

This leaves only d_3. Unfortunately, d_3 was the weakest of the three metrics considered. Presumably, the problem is that if a frame has little spectral contrast, random features will be blown up to have the same significance as the clear formant structure in a vowel. We have therefore introduced a modified contrast normalization, which can be written:

$$p' = \frac{cp}{\sigma + c} \qquad (1)$$

where p is a cepstrum coefficient and p' its normalized equivalent, σ is the standard deviation of the cepstrum coefficients in the frame, and c is a constant.

For frames in which $\sigma >> c$ the values are set such that the norm becomes c; whereas when $\sigma << c$ the values are unchanged. This prevents the enhancement of insignificant structure in low-contrast frames.

For $c = \infty$ the process reduces to the conventional non-normalized representation; while for $c = 0$ it becomes d_3. If the partial normalization represented by eqn. 1 is well motivated, an intermediate value of c will give the best results.

4. GENERAL EXPERIMENTAL DETAILS

All speech data used here was sampled at 8 kHz and analyzed by a 17-channel mel-scale filter-bank with a 16ms frame rate.

The thresholding applied to the log energies was uniform across the 17 channels. In experiments in trucks this absolute thresholding was augmented by a threshold applied at a given number of dB below the peak energy in the frame. This peak-related thresholding thus limits the dynamic range allowed across the channels in each frame.

It is not obvious how to apply spectral normalization to parameter sets including dynamic spectral information. Carlson and Clements appended their dynamic parameters to the static ones before computing the contrast over the extended set. After some preliminary experiments we decided to compute the contrast normalization over the C_1 to C_{16} static cepstral representation only and derive our dynamic parameters by taking three-frame differences between these normalized cepstra. C_0 and δC_0 were appended to the normalized parameters, giving a total of 34 parameters in all, which were then weighted in a manner equivalent to using grand variance weighting in the distance measure.

When an IMELDA representation was used, just 17 transformed parameters were derived from the 34 (possibly normalized) cepstrum parameters.

The recognition process used one HMM per word in the vocabulary. Each state was represented by a single centroid and the state-to-frame comparison used an unweighted Euclidean distance.

5. TESTS WITH ADDED WHITE NOISE

In these speaker-independent tests a downsampled subset of a studio-recorded alphabet database widely distributed in the U.K. was used. The training set consisted of one example of each letter from each of 26 male speakers, while the test set consisted of one example from each of 27 other male speakers.

The first set of tests compared the performance of IMELDA with that of weighted cepstrum representations when steady white noise was added to the test speech. The threshold was set at a level appropriate for noise-free speech (labeled arbitrarily 0 dB) and no contrast normalization was applied. The word models and the variances for the cepstrum representation were derived from the undegraded training data. The within-class covariance data for the IMELDA transformation was, however, derived from training data to which white noise had been added to 15 dB SNR, together with the same data filtered to apply a 6 dB/octave tilt as well as the undegraded data.

Table 1 shows the test results at various SNR levels. With no noise added, IMELDA has slightly better performance than the weighted cepstrum, even though it needs only half as many parameters, halving the computational cost of the Euclidean distance calculation.

In the conditions with added noise and spectral tilt the IMELDA representation shows a large performance advantage. The advantage at 15 dB SNR is consistent with that at the other noise levels, confirming that the noise resistance of IMELDA is not limited to the level presented in the derivation of the transformation.

Table 1. Error rates in speaker-independent alphabet recognition tests with low thresholds

test condition	IMELDA Transform	Weighted Cepstrum
undegraded	10.2%	12.4%
6 dB/octave tilt	13.5%	29.3%
20 dB SNR	30.4%	60.7%
15 dB SNR	44.8%	78.2%
10 dB SNR	61.0%	87.9%

Because of its strong advantage in noise, experiments on thresholding and contrast normalization were confined to the IMELDA representation. The effect on an IMELDA representation with the 0 dB threshold and partial contrast normalization was then explored. (A small difference in the way in which the IMELDA transform was derived made these results slightly different from those in Table 1.) Fig. 1 shows the results at 15 dB SNR, confirming the expectation that a finite, non-zero value of the constant c, around $c = 3$ here, would give best results. Also as expected, partial contrast normalization was not helpful on speech without added noise; but it was not harmful either, with the recognition rate at $c = 3$ being virtually identical to that with no normalization.

Fig. 1 also shows the effect of contrast normalization on speech at 15 dB SNR when the threshold level is optimized for this condition: namely, 20 dB higher than before. The performance at $c = 3$ and with no contrast normalization is similar.

Fig. 2 shows the performance as a function of the threshold level with speech at 15 dB SNR and without added noise. When partial contrast normalization is applied at $c = 3$, optimum performance in noise-free conditions is unchanged and that at 15 dB SNR is only slightly improved. However, with such contrast normalization, a relatively high level of recognition accuracy for the 15 dB SNR condition is maintained over a wider range of threshold levels: that is, to thresholds below the optimum level. In noise-free conditions, good performance may also be extended to slightly higher threshold levels.

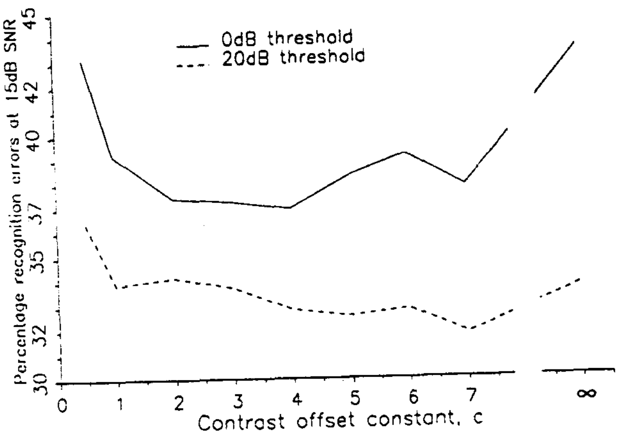

Fig. 1. Speaker-independent alphabet recognition error rates in white noise at 15 dB SNR as a function of the offset constant, c, in the contrast normalization. A value of infinity corresponds to no contrast normalization. Results are shown at a low spectral threshold value, 0 dB, and at 20 dB, the optimal value for this SNR.

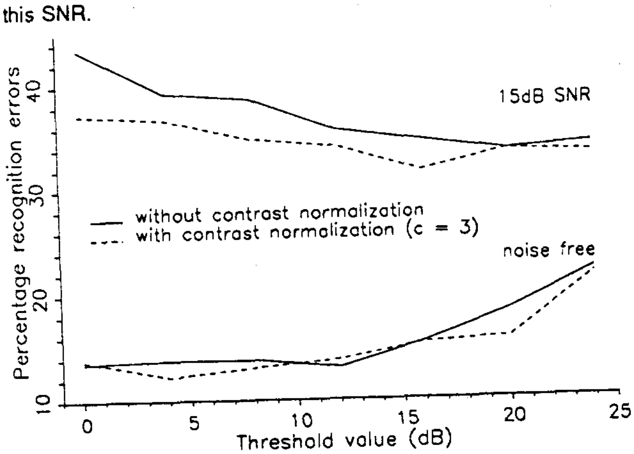

Fig. 2. Speaker-independent alphabet recognition error rates as a function of the energy threshold level. Results are shown with and without contrast normalization and in the 15 dB SNR and noise-free conditions.

6. TESTS WITH SPEECH IN TRUCKS

Isolated digits were recorded by three male passengers in both a light truck and a heavy truck during urban and fast highway driving. A non-directional Motorola microphone attached to the door pillar was used, resulting in test-speech recordings with low and rapidly varying SNRs. The reference speech was recorded in similar conditions, but with a boom-mounted, noise-canceling Shure SM-10 microphone as well as with the Motorola microphone. The SM-10 recordings were largely noise free, with SNRs around 20 dB greater than those of the test speech. Single word models were made from all conditions and the three speakers using the SM-10. IMELDA transforms were then derived by matching the simultaneously recorded Motorola recordings to these models.

The noise spectrum averaged over the Motorola recordings was found to be flat, suggesting that contrast normalization would be appropriate. With the energy threshold set low, Fig. 3 confirms that the recognition performance on the 375 test digits is helped by partial contrast normalization. Despite the wide variation in noise levels across the conditions, it was found that a single set of absolute and peak-related thresholds could bring the error rate down to zero, making further improvement through contrast normalization impossible. Disappointingly, there was also no evidence that contrast normalization broadened the range of threshold values over which good results could be obtained. No cepstrum tests were carried out on this data.

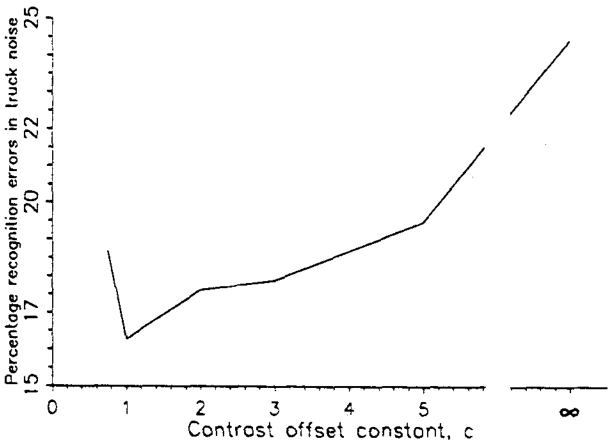

Fig. 3. Multi-speaker isolated-digit recognition error rate in moving trucks as a function of the offset constant, *c*, in the contrast normalization for an inappropriately low value of the energy threshold. An infinite value of *c* corresponds to no contrast normalization.

7. CONCLUSIONS

In noisy or distorting conditions an IMELDA spectra representation has been confirmed to give better result than a weighted cepstrum ~~representation~~, and ~~at lowe~~ computational cost. The tests with added white nois show that this advantage does not depend on th IMELDA derivation's having been exposed to data o the same noise level.

A partial contrast normalization technique has beer described, which is compatible with unweightec Euclidean distances and is therefore computationall efficient. This contrast normalization improved performance in both steady white noise and in the time varying broadband noise in the truck recordings, but onl when the thresholding on the log power spectrum wa set at an inappropriate level. In the case of the stead white noise, it also broadened the range of threshol values over which good performance could be obtained.

ACKNOWLEDGMENTS

We are grateful to British Telecom for providing the studio-quality alphabet recordings and to NRC Canada for allowing us to use some software. We thank Steven Soon for help with some of the experiments.

REFERENCES

1. D.H. Klatt, "A Digital Filter Bank for Spectral Matching," *Proc. IEEE Int Conf. Acoustics, Speech & Signal Processing, ICASSP-79*, Washington DC, April 1979, pp. 573-576.
2. Steven F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. on Acoustics, Speech & Signal Processing*, Vol. 27, No. 3, April. 1979. pp. 113-120.
3. M.J. Hunt and C. Lefebvre, "A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech," *Proc. IEEE Int Conf. Acoustics, Speech & Signal Processing, ICASSP-89*, Glasgow, Scotland, May 1989, pp. 262-265.
4. M.J. Hunt, D.C. Bateman, S.M. Richardson and A. Piau, "An Investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination," *Proc. IEEE Int Conf. Acoustics, Speech & Signal Processing, ICASSP-91*, Toronto, Canada, May 1991, pp. 881-884.
5. David Mansour and Biing Hwang Juang, "A Family of Distortion Measures Based Upon Projection Operation for Robust Speech Recognition," *IEEE Trans. on Acoustics, Speech & Signal Processing*, Vol. 37, No. 11, Nov. 1989, pp. 1659-1671.
6. Beth A. Carlson and Mark. A. Clements, "Application of a Weighted Projection Measure for Robust Hidden Markov Model Based Speech Recognition," *Proc. IEEE Int Conf. Acoustics, Speech & Signal Processing, ICASSP-91*, Toronto, Canada, May 1991, pp. 921-924.

A ROBUST CONNECTED-WORDS RECOGNIZER

S. Dobler, P. Meyer, H.W. Ruehl

Philips Kommunikations Industrie AG
Thurn-und-Taxis-Str. 14
D-8500 Nuernberg

ABSTRACT

In the context of speaker independent recognition of connected words, this paper addresses several problems:

- Robustness against variations of both background noise and frequency responses of microphones and transmission lines is improved. This is achieved employing a high-pass filter to reduce stationary or slowly varying parts of the spectral components in the feature vectors.
- A method is proposed to improve the power of references generated from isolated words by combining them with duration models derived from a small set of connected words.
- Experiments to improve recognition accuracy using speaker dependent adaptation of the HMM models' transition probabilities and of feature vectors are presented.

Results are given for a German corpus and the TI/NIST connected digits corpus.

1. INTRODUCTION

Currently, a standard approach to achieve high-accuracy speaker independent recognition is to increase feature vector length using not only spectral or cepstral components, but also their first and second order time derivatives, and additionally use mixture density HMM state modelling with up to 128 mixtures per state [1, 2]. For centralized applications with small vocabularies, e.g. for telephone service automation, this is an adequate approach, as performance and not costs need to be optimized. But for terminal applications like e.g. voice control for telephones or consumer electronics devices, costs are an important factor favouring low-complexity algorithms.

The mixture density approach is well suited to model both dialectal pronounciation variants for speaker independency, and coarticulation variants at word boundaries for utterances containing connected words. But up to now, no solution exists to introduce speaker adaptivity to HMM mixture-density models, as it is impossible to separate dialectal and coarticulatory effects in the models. Such a separability may be of advantage in dialogue-intensive applications (e.g. mail ordering) for languages with several dialectal variants. In such cases, confusions may be rather unlikely, if the references can be focused to the dialect actually used, but very likely, if all dialectal variants have to be considered.

For these two reasons, we experimented with a 'small' recognition algorithm employing only single HMM densities. Our main goals were to find short, robust feature vectors, and to examine speaker independence in combination with speaker specific adaptation of both densities and transition probabilities.

2. RECOGNITION ALGORITHM

The recognizer uses hidden Markov modelling (HMM) for training, generating two templates (a male and a female) for each word, with single continuous Laplacian densities and unquantized observations. The number of states of the HMM models is determined by the average length of the corresponding training utterances, using identical transition probabilities at all states of the models. Fig. 1 sketches the HMM models. Both training of the HMM models and recognition use the Viterbi algorithm.

To calculate feature vectors, speech is sampled at 8 KHz, weighted by a Hamming window of 32 ms length and FFT-transformed. Adjacent windows overlap by 20 ms. The FFT-based power density spectrum is smoothed and down-sampled to 15 components equally spaced on the mel scale. Each component is replaced by its logarithm and normalized to the average energy of the feature vector, with average

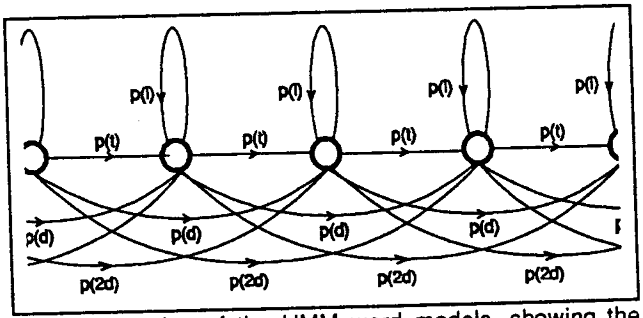

Fig 1: A section of the HMM word models, showing the identical transition probabilities for all transition types

0-7803-0532-9/92 $3.00 © 1992 IEEE

energy being added to the feature vector as 16th component. More details about the recognition algorithms may be found in [3].

3. INCREASING ROBUSTNESS

To evaluate the recognizer's speaker independence and its robustness against environmental changes or changes of background noise, a corpus MMIX for German digits was used. It consists of several subcorpora collected at five locations all over Germany and Austria in order to cover dialectal pronounciation variants. MMIX contains about 3000 digits spoken in isolation, and about 340 strings of three and 380 strings of seven digits spoken by about 80 male speakers. Both versions of 2 ('zwo', and 'zwei' confusable to 'drei' = 3) are used.

Collecting environment ranged from unechoic chambers via conference rooms to noisy office rooms. Different microphones and recording facilities were used containing handheld microphones, telephone handsets, speaker phones, and in some cases transmission via local or long distance telephone lines. Average signal-to-noise ratio (SNR) was better than 30 dB, covering a range from 25 to 40 dB.

As the number of multiple-digit strings was rather limited, only isolated digits were used for training, saving all digit strings for the evaluation. 26 speakers of MMIX were selected for training in such a way that all dialects and recording conditions were covered representatively. Each speaker uttered each digit three to four times, resulting in 60 to 80 training utterances per digit. This turned out to be a sufficiently large training set, as further increasing the training data size did not improve recognition accuracy. Furthermore, no statistically significant difference was measured for recognition results on training and test data set.

To control the influence of the inhomogeneity of the corpus, a sub-corpus MPFH containing isolated digits spoken in a quiet room by 56 male speakers from one region was evaluated separately.

corpus / digit string length	without	with			
	high-pass filtering				
	errors / %	subs	del	ins	errors / %
MPFH / 1	1.41	7	2	9	0.97
MMIX / 1	19.07	14	13	12	0.96
MMIX / 3	11.53	9	1	0	0.97
MMIX / 7	21.82	23	34	1	2.04

Tab. 1: Digit errors for speaker independent recognition with unknown string length

Using the homogeneous corpus MPFH both for training and testing, an acceptable digit error rate of 1.41% was achieved. But for the inhomogeneous corpus MMIX, digit errors rose to more than 10% (see Tab. 1). This increased error rate is mainly caused by the fact that single-density HMMs can model speech only for constant conditions, i.e. for one kind of transfer function of the recording facility, and only for stationary background noise. Mixture density HMMs are less sensitive to these effects, as they can use their vector pools to model environmental changes, too, at the cost of significantly increased computational effort compared to single density HMMs. For this reason, our approach was not to model different noise and transfer function conditions, but to find means to reduce their influence from the signal representation.

As most of the unwanted information is stationary and thus producing offsets for the spectral components on either the linear (for background noise) or the logarithmic (for transfer functions) scale, several kinds of modulation frequency high-pass filtering, among them e.g. delta spectrum, were examined. More information concerning these experiments may be found in [4]. It turned out that filtering both the log. scaled energy and the log. scaled spectal components over time using a simple first order IIR high-pass of the form

$$c_{new}(n) = c_{old}(n) - c_{old}(n-1) + 0.7\ c_{new}(n-1)$$

with a cut-off frequency of approximately 4.5 Hz improved recognition most. As shown in Tab. 1, digit error rates for MMIX are reduced by an order of magnitude to 1% resp. 2% for the 7-digit strings. On the homogeneous control data base MPFH, the improvement caused by high-pass filtering is only from 1,4% to 1%, showing that high-pass filtering is of little use to improve speech recognition under controlled conditions, but is very effective to reduce variance by environmental effects.

To find out what happens if the training conditions are significantly different from the recognition conditions, the speaker independent models generated from MMIX were tested against a database AUTO containing digit strings from 8 male speakers collected via handset in a moving car at 120 km/h, with a background noise level of approximately 80 dB. Digit error rates are given in Tab. 2.

For the single digits, the SNR is about 5 dB below the worst training cases. But still, recognition works excellent. SNR has to be decreased by more than 12 dB compared to training to increase digit error rate to 3.7%. For comparison, without high-pass filtering, recognition on the AUTO corpus is

string length	av. SNR / dB	errors / %
AUTO / 1	19	0.6
AUTO / 3	13	3.7
AUTO / 7	10	5.3

Tab. 2: Digit error rates for speaker independent recognition on the corpus AUTO collected in a car

impossible due to more than 100% insertions. So, high-pass filtering not only improves robustness against variations covered by training data, but also against variations beyond the training data range.

To our opinion, the efficiency of high-pass filtering is caused by three effects:

The frequency responses of recording environment and transmission lines are cancelled, due to the fact that logarithmically scaled spectral components are filtered. In the logarithmic domain, frequency responses are converted into a constant offset of the spectral components, being removed completely by high-pass filtering.

In speech pauses and during low-energy parts of speech, effects caused by stationary or slowly changing background noise are reduced. This kind of background noise causes a constant offset in the linear domain. After taking the logarithm, it has neglegible effect on the high-energy parts of the speech. But it produces an almost constant non-zero component level, if the speech level is below the background noise level. As this level changes with background noise level, increased background noise usually results in lots of insertions. High-pass filtering reduces the constant level to zero and thus inhibits insertions in speech pauses.

At last, the very speaker dependent low-frequency parts of the modulation spectrum are removed. It is known from speaker independent isolated-words recognition [5], that normalisation of the spectral feature vectors with respect to the long-term spectrum improves recognition accuracy. High-pass filtering replaces long-term spectrum by a gliding short-term spectrum which is not quite as effective as a long-term spectrum, but still has benefits.

In the following sections, the TI/NIST corpus will be used for further experiments. The recognizer as described in this section is used as a baseline system with a string error rate of 5.8% as given in Tab. 3, achieved with embedded training using the complete specified training data set. This corresponds roughly to a digit error rate of 1.8% which is rather high in comparison to 2% for German 7-digit strings achieved with isolated word training. Exept for the different vocabularies, the main reason for this difference seems to be that the TI digit strings are spoken rather coarticulated, whereas most German strings are pronounced in a clear and distinct fashion.

In [7], a very similar recognizer is described using the same recognition algorithm and HMM models with one reference per digit per gender, but using a feature extraction after [8]. Its string error rates of 8.6% for Laplacian densities resp. 7.2% for Gaussian densities correspond to 5.8% achieved by our recognizer, showing that the 16 HP-filtered spectral components perform better than 12 LPC-cepstral and 12 delta-cepstral coefficients plus one log-energy and one delta log-energy component.

4. POSITION DEPENDENT DURATION MODELLING

As our corpus of German connected digit utterances still is limited, we have been looking for ways to improve our references created from isolated words. A preliminary experiment proved that better modelling of duration by modifying transition probabilities significantly reduced the error rate. Closer examination of spoken German digit strings showed that a digit's length is strongly influenced by its position in a phrase, i.e. whether the digit is spoken isolated, or in initial, medial, or final position of an utterance.

To find out how much HMM models with position dependent length improve recognition accuracy, the TI/NIST digit string corpus [6] was used to train references according to a syntax model given in Fig. 2, with a male and a female reference template per digit per arc. This syntax model was developed for the German language, where it nicely covers positional duration variations, but seems to be somewhat oversized for American English. In the TI/NIST corpus, only prepausal lengthening effects were observed.

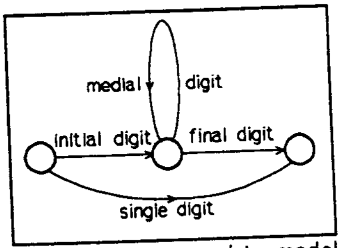

Fig. 2: Syntax used to model position dependent digit durations

Results are given in Tab. 3. Using the syntax after Fig. 2 to train position dependent references, and for evaluation, the string error rate is reduced from 5.8% to 4.6%.

On the other hand, using only references generated from isolated digits, string error rate increases to 10.3%. To reduce this error rate in cases when sufficient data is only available for isolated-word training, initial, medial and final

	str.error rate / %
/4/, Gaussian densities	7.2
/4/, Laplacian densities	8.6
1 optimal reference / word / gender	5.8
4 opt. position dependent references	4.6
1 ref., trained only with single digits	10.3
4 ref. with position dependent duration	6.6
1 ref. + adapt. transition probabilities	5.2
1 ref. + adapt. feature vectors	4.7
1 ref. + adapt. transition probabilities + adapt. feature vectors	4.1

Tab. 3: String error rates for the TI/NIST corpus

references were aligned with their corresponding single-word references, and their feature vectors were replaced by the closest-match feature vectors of the single-digit references. The combination of position dependent duration models and feature vectors generated from isolated words decreased string error rate to 6.6%.

This result indicates that about half of the advantage of embedded training over isolated-word training is gained by better duration modelling of the references. The next step for duration modelling will be to explore how to create reliable duration models using only a small subset of the training digit strings. This will allow to quickly generate improved references for connected-words recognition, based on existing training data spoken in isolated fashion, with only a small connected-words training corpus to be newly collected.

5. SPEAKER SPECIFIC ADAPTATION OF DURATIONS AND FEATURE VECTORS

The following experiments concerning speaker adaptation were done with no syntax and one reference per digit per gender. The test database was ordered in such a way that all utterances of a speaker were evaluated in sequence, starting with the isolated digits, the other utterances following in ascending string length. Adaptation is done using only utterances that have already been recognized, but in a supervised way, i.e. using the correct result and not the recognition result.

The idea behind this procedure is that in applications requiring a large amount of input data like e.g. mail ordering, users will accept a start phase with rather slow data entry and potential corrections to let the recognizer adapt to their voice, but they want increased speed of data entry after the start phase.

Whereas position dependent duration models relate to language specific features, this experiment tried to model speaker specific variations of duration, especially speaker dependent average talking speed. In the HMM models used, all states have identical transition probabilities, due to the variable number of states per model. This fact allows easy adaptation without supervision by simply changing transition probabilities during recogition according to the recent number of state skips and repeats.

Rules to adapt transition probabilities to talking speed were derived in the following way: HMM models were generated using the standard training. Then, the training data set was subdivided into several classes with different average word length. For each class, the transition probabilities of the HMM models were retrained thus getting optimal probabilities for the word length class. From the set of probabilities for each transition class, a curve was derived using linear regression, allowing calculation of optimal transition probabilities for any given word length.

Using the regression lines to adapt transition probabilities after each test string, employing one reference per digit per gender, and no syntax, string error rate was reduced from the baseline 5.8% to 5.2%.

As the TI/NIST database covers a wide range of regional accents, it was used to examine speaker specific adaptation of feature vectors, too.

Starting with speaker independent references for every new speaker, the isolated digits were first recognized, and immediately after recognition used to adapt feature vectors of the correct reference model to the speaker specific pronunciation. No adaptation was done using the multiple-digit strings.

For adaptation, speaker specific vectors got a weight of a quarter of the reference vector that they were mapped to. As there were two utterances of each isolated digit available per speaker, the digit strings were recognized with adapted references containing 36% speaker specific information plus 64% original speaker independent information per feature vector. Feature vector adaptation reduced string error rate from 5.8% to 4.7%.

Combining adaptive transition probabilities and feature vector adaptation further reduced error rate to 4.1%.

REFERENCES

[1] H. Ney: "Acoustic-Phonetic Modeling Using Continuous Mixture Densities for the 991-Word DARPA Speech Recognition Task", Proc. ICASSP 90, pp. 713-716, Albuquerque, 1990

[2] J.G. Wilpon, C.-H. Lee, L.R. Rabiner: "Improvements in Connected Digit Recognition Using Higher Order Spectral and Energy Values", Proc. ICASSP 91, pp. 349-352, Toronto, 1990

[3] H.W. Ruehl et al.: "Speech Recognition in the Noisy Car Environment", Speech Communication, vol. 10, no. 1, 1991, pp. 11-22

[4] H.G. Hirsch, P. Meyer, H.W. Ruehl: "Improved Speech Recognition Using High-Pass Filtering of Subband Envelopes", Proc. EUROSPEECH 91, Genova, 1991

[5] M.H. Kuhn, H.H. Tomaschewski: "Improvements in Isolated Word Recognition", IEEE Transact. ASSP, Vol. ASSP-31, No. 1, pp. 157-167, 1983

[6] R.G. Leonhard: "A Database for Speaker Independent Digit Recognition", Proc. ICASSP 84, San Diego,CA., 1984, pp. 42.11.1-4

[7] H. Ney: "Speech Recognition in an Neural Network Framework: Discriminative Training of Gaussian Models and Mixture Densities as Radial Basis Functions", Proc. ICASSP 91, Toronto, 1991, pp. 573-576

[8] L.R. Rabiner, C.H. Lee, B.H. Juang, J.G. Wilpon: "HMM Clustering for Connected Word Recognition", Proc. ICASSP 89, Glasgow, pp. 405-408, 1989