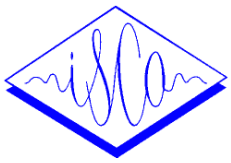

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A COMPARISON OF THREE NOISY SPEECH RECOGNITION APPROACHES

Olivier Siohan *Yifan Gong* *Jean-Paul Haton*
CRIN–CNRS & INRIA Lorraine, BP 239, 54506 Vandoeuvre–lès–Nancy, France

ABSTRACT

We compare 3 recent approaches dealing with speech recognition in noisy environment. The first approach is based on stochastic model combination of noise and speech . Given a clean speech model based on speech trajectories and an HMM noise model, this method aims at deriving a noisy speech model, in order to recognise noisy speech. In the second approach, we perform a mapping between the noisy and the clean speech space. The noisy speech is recognised after mapping to the clean space, using clean speech models. In the last approach, LDA is used as a preprocessing, and the training and testing environmental conditions are identical. On a 206 isolated word recognition task under different noisy environment, LDA gave the best results. The model combination proved to be efficient at high SNR, but performances fell down at low SNR. The mapping approach showed to be very robust, but led to the lowest recognition rate at high SNR.

1. INTRODUCTION

In this paper, we compare 3 recent approaches for dealing with noisy speech recognition: a stochastic model combination, a parameter space mapping procedure called base transformation and a robust speech parametrisation based on Linear Discriminant Analysis.

The first approach assumes that clean speech is modelled by a set of Stochastic Trajectory Models (STM) [1] and the background noise by an ergodic HMM. Given clean speech STMs and the noise HMM, we compute a set of STMs in order to recognise noisy speech, based on the assumption that speech and noise are uncorrelated and additive in the linear power spectral domain. In this domain, a noisy observation can be viewed as the combination of a clean speech observation and a noise observation. The combination of the probability density functions (pdf) is performed using a similar approach as Gales [2], where noisy STM parameters are derived from clean STM parameters and noise HMM parameters.

The base transformation approach [3] assumes that recognition performance in noisy environment may be improved by transforming the noisy speech into a reference environment, and then recognising it in the reference environment. The environmental difference is converted into a base difference and removed by a base transformation. An environment is defined by a base. A vector is expressed as a linear combination of base elements. Each base element has a phonetic label, and the equivalence between the training and the operating bases is obtained by an automatic alignment of the same text uttered in the two environments. The transformation of a noisy vector consists in changing the base while keeping the same linear combination of base elements.

In the last approach, training and test are performed in similar environmental conditions, but the signal is represented using a LDA transformation. In the original parameter space, noisy speech is represented by MFCC and ΔMFCC parameters. Using LDA, we project this space into a smaller dimensional subspace, where the between class variance is maximised and the within class variance is minimised.

The paper is organised as follows. Section 2, 3 and 4 respectively present the framework of stochastic model combination, the base transformation approach and the LDA preprocessing. Section 5 describes the experiments and results, and section 6 concludes the paper.

2. STOCHASTIC MODELS COMBINATION

2.1. Noise and speech pdf combination

Let $x(t)$ be the clean speech signal and $n(t)$ the noise signal in the time domain. Under additivity assumption, and assuming the noise does not alter the frame location, the noisy speech signal $y(t)$ is:

$$y(t) = x(t) + n(t) \tag{1}$$

In the power spectral domain, and under noise and speech independence assumption, Eq-1 is rewritten as:

$$Y(\omega) = X(\omega) + N(\omega) \tag{2}$$

where $Y(\omega)$, $X(\omega)$ and $N(\omega)$ are respectively the power spectral densities of $y(t)$, $x(t)$ and $n(t)$.

Let $\mathbf{Y}^{cep}$, $\mathbf{X}^{cep}$ and $\mathbf{N}^{cep}$ be the cepstrum vectors of $y(t)$, $x(t)$ and $n(t)$. In the cepstral domain, $\mathbf{Y}^{cep}$, $\mathbf{X}^{cep}$ and $\mathbf{N}^{cep}$ are random vectors. We suppose $\mathbf{X}^{cep}$ and $\mathbf{N}^{cep}$ are modelled as Gaussian distributions with means $\mathbf{m}^{cep}$ and $\tilde{\mathbf{m}}^{cep}$ and covariance matrices $\mathbf{\Sigma}^{cep}$ and $\tilde{\mathbf{\Sigma}}^{cep}$. Under such an assumption, and using the combination rule between speech and noise, our aim is to derive the expression of the noisy speech pdf.

Let $\mathbf{X}^{log}$ be the Discrete Cosine Transform of the random vector $\mathbf{X}^{cep}$. $\mathbf{X}^{log}$ characterises the power spectral density of speech in the log-spectral domain. If $\mathbf{W}$ is the Discrete Cosine Transform matrix, we have:

$$\mathbf{X}^{log} = \mathbf{W}\mathbf{X}^{cep} \tag{3}$$

Due to the linear relation between $\mathbf{X}^{log}$ and $\mathbf{X}^{cep}$, if $\mathbf{X}^{cep}$ is modelled as a Normal distribution with mean $\mathbf{m}^{cep}$ and covariance $\mathbf{\Sigma}^{cep}$, then the pdf of $\mathbf{X}^{log}$ is Normal, with mean $\mathbf{m}^{log}$ and covariance $\mathbf{\Sigma}^{log}$:

$$\mathbf{m}^{log} = \mathbf{W}\mathbf{m}^{cep} \tag{4}$$

$$\mathbf{\Sigma}^{log} = \mathbf{W}\mathbf{\Sigma}^{cep}\mathbf{W}^{t} \tag{5}$$

Let $\mathbf{X}^{lin}$ be the random vector which characterises the power spectral density of speech in the linear spectral domain:

$$\mathbf{X}^{lin} = \exp(\mathbf{X}^{log}) \tag{6}$$

If $\mathbf{X}^{log}$ is modelled as a Normal distribution with mean $\mathbf{m}^{log}$ and covariance $\mathbf{\Sigma}^{log}$, then $\mathbf{X}^{lin}$ is modelled as a

Log-Normal distribution, with mean $\mathbf{m}^{lin}$ and covariance $\mathbf{\Sigma}^{lin}$ [2]:

$$m_i^{lin} = \exp(m_i^{log} + \Sigma_{i,i}^{log}/2) \quad (7)$$

$$\Sigma_{i,j}^{lin} = m_i^{lin} m_j^{lin} [\exp(\Sigma_{i,j}^{log}) - 1] \quad (8)$$

Similar expressions can be derived for the noise vector $\mathbf{N}^{lin}$ in the linear spectral domain.

At this point, the Log-Normal pdfs of $\mathbf{X}^{lin}$ and $\mathbf{N}^{lin}$ are known. According to Eq-2, we have $\mathbf{Y}^{lin} = \mathbf{X}^{lin} + \mathbf{N}^{lin}$. The noisy speech pdf in the linear spectral domain is the convolution product of $\mathbf{X}^{lin}$ pdf and $\mathbf{N}^{lin}$ pdf. In order to simplify the expression of $\mathbf{Y}^{lin}$ pdf, we assume that this pdf is Log-Normal, with mean $\widehat{\mathbf{m}}^{lin}$ and covariance matrix $\widehat{\mathbf{\Sigma}}^{lin}$. Due to the independence assumption between noise and speech, we have:

$$\widehat{\mathbf{m}}^{lin} = g\mathbf{m}^{lin} + \tilde{\mathbf{m}}^{lin} \quad (9)$$

$$\widehat{\mathbf{\Sigma}}^{lin} = g^2\mathbf{\Sigma}^{lin} + \tilde{\mathbf{\Sigma}}^{lin} \quad (10)$$

We introduce a g correction factor, to deal with differences in speech energy level between training and testing conditions. Given the $\mathbf{Y}^{lin}$ pdf, the derivation of the parameters of $\mathbf{Y}^{log}$ pdf and $\mathbf{Y}^{cep}$ pdf is straightforward, by inverting Eq-7, Eq-8, Eq-4 and Eq-5. Finally, in the cepstral domain, the $\mathbf{Y}^{cep}$ pdf is a Normal distribution, with mean $\widehat{\mathbf{m}}^{cep}$ and covariance $\widehat{\mathbf{\Sigma}}^{cep}$. Both $\widehat{\mathbf{m}}^{cep}$ and $\widehat{\mathbf{\Sigma}}^{cep}$ are function of $\mathbf{m}^{cep}$, $\tilde{\mathbf{m}}^{cep}$, $\mathbf{\Sigma}^{cep}$ and $\tilde{\mathbf{\Sigma}}^{cep}$. Let $\theta = (\mathbf{m}^{cep}, \mathbf{\Sigma}^{cep})$ and $\tilde{\theta} = (\tilde{\mathbf{m}}^{cep}, \tilde{\mathbf{\Sigma}}^{cep})$, we have:

$$\widehat{\mathbf{m}}^{cep} = f_m(\theta, \tilde{\theta}) \;\; ; \;\; \widehat{\mathbf{\Sigma}}^{cep} = f_\Sigma(\theta, \tilde{\theta}) \quad (11)$$

2.2. Speech Stochastic Trajectory Model

In a parametric space, speech is represented by a point which moves as articulatory configuration changes. We define a trajectory of speech as a sequence of moving points.

Let $\mathbf{X}$ be a trajectory of Q fixed prespecified points: $\mathbf{X} = (\mathbf{x}_0, \ldots, \mathbf{x}_{Q-1})$, where each point is a D-dimensional vector in a parameter space. $\mathbf{X}$ is assumed to be obtained by re-sampling a sequence of d frames according to a linear time scaling. In our formulation, each phoneme symbol is associated with a set of stochastic generators of trajectories. A phone model may be viewed as a mixture of trajectory models. We defined the probability of sequence $\mathbf{X}$, given the phoneme symbol s and the duration d as [1]:

$$p(\mathbf{X}|d,s) \triangleq \sum_k Pr(T_k|s) \cdot p(\mathbf{X}|T_k,d,s) \quad (12)$$

where

- $p(\mathbf{X}|T_k,d,s)$ is the pdf of the vector sequence $\mathbf{X}$ given the trajectory component T_k, the duration d and the phoneme s,
- $Pr(T_k|s)$ is the probability of the component trajectory T_k given the phoneme s, with $\forall s, \; \sum_k Pr(T_k|s) = 1$.

Assuming that each of the Q points of the component trajectory T_k is produced by an independent distribution, the pdf of $\mathbf{X}_n$ given T_k, d and s is modelled as a multivariate Gaussian distribution, with mean $\mathbf{m}_{k,i}^s$ and covariance matrix $\mathbf{\Sigma}_{k,i}^s$:

$$p(\mathbf{X}_n|T_k,d,s) \triangleq \prod_{i=0}^{Q-1} \mathcal{N}(\mathbf{x}_{n-\frac{d}{2}+i\frac{d-1}{Q-1}}; \mathbf{m}_{k,i}^s, \mathbf{\Sigma}_{k,i}^s) \quad (13)$$

Suppose that noise is modelled at a given time as a Gaussian distribution with mean $\tilde{\mathbf{m}}$ and covariance matrix $\tilde{\mathbf{\Sigma}}$. The pdf of the noisy vector sequence is obtained combining Eq-13, Eq-1 and Eq-2:

$$p(\mathbf{Y}_n|T_k,d,s) \triangleq \prod_{i=0}^{Q-1} \mathcal{N}(\mathbf{y}_{n-\frac{d}{2}+i\frac{d-1}{Q-1}}; f_m(\theta_{k,i}^s, \tilde{\theta}), f_\Sigma(\theta_{k,i}^s, \tilde{\theta})) \quad (14)$$

2.3. Noise HMM

We use an ergodic HMM as a non stationary noise model. The HMM is characterised as follows:

- N is the number of states in the model,
- $p(q_{t+1} = j|q_t = l)$ is the state transition probability distribution, from state l to state j between time t and $t+1$,
- π_j is the initial state distribution,
- $\mathcal{N}(\mathbf{n}; \tilde{\mathbf{m}}_j, \tilde{\mathbf{\Sigma}}_j)$ is the observation symbol probability density function in state j.

2.4. Decoding stage

When the HMM state number N is greater than 1, we cannot directly apply Eq-14. We have to choose which HMM state to combine with a given STM state and mixture component.

Let $\delta_0(j)$ be the probability of the vector taken from the state 0 of the noisy trajectory $\mathbf{Y}_n$, for symbol s and duration d, when the noise HMM is in state j. We have:

$$\delta_0(j) = \pi_j \cdot \mathcal{N}(\mathbf{y}_{n-\frac{d}{2}}; f_m(\theta_{k,0}^s, \tilde{\theta}_j), f_\Sigma(\theta_{k,0}^s, \tilde{\theta}_j)) \quad (15)$$

Let $\delta_i(j)$, the probability of the vector taken from the state i of the noisy trajectory $\mathbf{Y}_n$, for symbol s and duration d, when the noise HMM is in state j. Using a Viterbi recursion, we have:

$$\delta_i(j) = \max_{l=1}^{N} \Big[\delta_{i-1}(l) \cdot p(q_{i\frac{d-1}{Q-1}} = j|q_{(i-1)\frac{d-1}{Q-1}} = l) \cdot \mathcal{N}(\mathbf{y}_{n-\frac{d}{2}+i\frac{d-1}{Q-1}}; f_m(\theta_{k,i}^s, \tilde{\theta}_j), f_\Sigma(\theta_{k,i}^s, \tilde{\theta}_j))\Big] \quad (16)$$

Finally, we have:

$$p(\mathbf{Y}_n|T_k,d,s) = \max_{l=1}^{N} \delta_{(Q-1)}(l) \quad (17)$$

Eq-17 is used instead of Eq-14 when the number of states in the HMM is greater than 1.

3. BASE TRANSFORMATION

3.1. Basic assumption

Let χ and ψ be two D-dimensional parameter spaces characterising two acoustical environments. Let ϕ be a set of J labels: $\phi = \{q_1, q_2, ..., q_J\}$. An observation is the acoustic representation of a label in a D-dimensional parameter space. Let $\mathbf{b}_i^\chi$ and $\mathbf{b}_i^\psi$ be the observation of label q_i respectively in χ and ψ. The set of all label observations constitutes a base. We have therefore two bases $\mathbf{B}^\psi = \{\mathbf{b}_i^\psi\}$ and $\mathbf{B}^\chi = \{\mathbf{b}_i^\chi\}$. Let $f : \chi \rightarrow \psi$ be a parameter mapping from one environment to the other. f is the base transformation, and we have: $\mathbf{b}_i^\psi = f(\mathbf{b}_i^\chi)$.

We suppose that

$$\forall \mathbf{v}^\chi \in \chi \;\; \exists \Omega_{\mathbf{v}}^\chi \subseteq \mathbf{B}^\chi$$

such that for vector $\mathbf{v}^\chi$, $\mathbf{v}^\psi = f(\mathbf{v}^\chi) \in \psi$, and $\forall \mathbf{b}_i \in \Omega_{\mathbf{v}}^\chi$:

$$\mu(\mathbf{v}^\chi, \mathbf{b}_i^\chi) = \gamma \cdot \mu(\mathbf{v}^\psi, \mathbf{b}_i^\psi) \quad (18)$$

where $\mu(\mathbf{a}, \mathbf{b})$ is a measure of the similarity between vectors $\mathbf{a}, \mathbf{b}$, and γ is a scaling factor constant for all vectors.

The mapping from χ to ψ preserves relative similarities of speech vectors to a subset of label observations, which are environment independent. In general, $\Omega_{\mathbf{V}}^{\chi}$ is made of the label observations that are closest to $\mathbf{v}^{\chi}$. $\Omega_{\mathbf{V}}^{\psi}$ is known since it is the observation of the same set of labels with $\Omega_{\mathbf{V}}^{\chi}$.

3.2. Adaptation by base transformation

Suppose the recognition system is trained with data from the ψ environment. In recognition, utterances from the χ environment are transformed frame by frame, using the mapping from the χ environment to the ψ environment. The transformation of $\mathbf{x}$ from the χ environment is performed in two steps. First, we compute the set of the μ values relating to different labels in the base $\mathbf{B}^{\chi}$. Then, an inverse transform is applied on the μ's with the base of ψ, $\mathbf{B}^{\psi}$. The recognition is performed on the inversely transformed utterances ($\chi \rightarrow \psi$) using a speech recognition system trained for the ψ environment.

We use the phonemes of a set of texts as labels. The observations of the labels are obtained from continuous pronunciation of the texts in the two environments. The utterance segments from ψ are labelled, and the utterance segments from χ are time aligned with those from ψ in order to obtain labelled speech for the two environments $\{(\mathbf{b}_i^{\chi}, \mathbf{b}_i^{\psi})\}$.

3.3. Transformation

For each input vector $\mathbf{x}_n$, $0 \leqslant n < N$, where N is the number of frames in the utterance, we use Eq-19 as the transform:

$$\forall \mathbf{b}_i^{\chi} \in \Omega_{\mathbf{x}_n}^{\chi} \quad \mu_i({\mathbf{x}_n}^{\chi}, \mathbf{b}_i^{\chi}) = \frac{s({\mathbf{x}_n}^{\chi}, \mathbf{b}_i^{\chi})^{\alpha}}{\sum_j s({\mathbf{x}_n}^{\chi}, \mathbf{b}_j^{\chi})^{\alpha}} \tag{19}$$

where $\alpha = \frac{1}{(m-1)}$, $m \in (1, \infty)$ and $s(x, y)$ is a similarity measure between the D-dimensional vectors x and y.

$$s(x, y) = \frac{1}{\sqrt{\sum_{k=1}^{D} \lambda_k (x_k - y_k)^2}} \in [0, \infty)$$

is based on the Euclidean distance weighted by λ_k. Eq-20 gives the inverse transform:

$${\mathbf{x}_n}^{\psi} = \frac{\sum_j (\mu_j)^m \cdot \mathbf{b}_j^{\psi}}{\sum_j (\mu_j)^m} \tag{20}$$

In Eq-19, 20, the summations on j are performed over all j's that satisfy $\mathbf{b}_j^{\chi} \in \Omega_{\mathbf{x}_n}^{\chi}$. $\Omega_{\mathbf{x}_n}^{\chi}$ can be chosen as the set of $K (K < J)$ label observations closest to $\mathbf{x}^{\chi}$.

The inverse transform is a linear combination of $\mathbf{b}_j^{\psi}$. Note that in the above formulation each base element has a phonetic label. That differs from traditional fuzzy vector quantisation where clusters are specified by only a distance metric and therefore have no direct phonetic interpretation.

4. LINEAR DISCRIMINANT ANALYSIS

LDA aims at improving discrimination between classes in a vector space, by finding a linear transformation from a D-dimensional vector space to a d-dimensional vector space ($d \leqslant D$). This transformation is defined according to the widely used criterion which maximises $\mathrm{tr}(\mathbf{W}^{-1}\mathbf{B})$, where $\mathrm{tr}(\mathbf{X})$ denotes the trace of matrix $\mathbf{X}$. $\mathbf{W}$ and $\mathbf{B}$ are the within and between class covariance matrices computed as:

$$\mathbf{B} = \frac{1}{N} \sum_{k=1}^{K} n_k (\mu_k - \mu)(\mu_k - \mu)^t \tag{21}$$

$$\mathbf{W} = \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{n_k} (x_{kn} - \mu_k)(x_{kn} - \mu_k)^t \tag{22}$$

where N is the total number of training pattern, K is the number of class, n_k is the number of training pattern of the kth class, and μ_k and μ are respectively the mean of the kth class and the overall mean, given by:

$$\mu_k = \frac{1}{n_k} \sum_{n=1}^{n_k} x_{kn} \quad ; \quad \mu = \frac{1}{N} \sum_{k=1}^{K} n_k \mu_k \tag{23}$$

where x_{kn} is the nth training pattern from the kth class.

Using the optimisation criterion defined above, it can be shown that the d column vectors of the transformation matrix $\mathbf{U}$ are the d eigenvectors associated to the d largest eigenvalues of the matrix $\mathbf{W}^{-1}\mathbf{B}$. The computation of eigenvectors and eigenvalues are performed using the method described in [4]. Let $\mathbf{C}$ be the unitary matrix diagonalizing $\mathbf{W}$ to $\mathbf{L}$, $\mathbf{W} = \mathbf{CLC}^t$. Let $\mathbf{V}$ be the unitary matrix whose column vectors are the d eigenvectors corresponding to the d largest eigenvalues of the symmetric matrix $\mathbf{S} = \mathbf{L}^{-1/2}\mathbf{C}^t\mathbf{BCL}^{1/2}$. Then, the matrix $\mathbf{U}$ is given by $\mathbf{U} = \mathbf{CL}^{-1/2}\mathbf{V}$.

In our formulation, a class is associated to a phoneme. The matrix computation is done using a training corpus labelled at phonetic level. We use the matrix multiplication as a preprocessing step in the front-end. The test corpus is transformed using the matrix computed on the training corpus.

5. EXPERIMENTS AND RESULTS

5.1. Experimental settings

Experiments were performed on an isolated words recognition task in a speaker dependent mode, using our VINICS speech recognition system, based on Stochastic Trajectory Models [1]. Speech units are context independent phonemes.

We used 130 sentences, randomly selected from the TIMIT database and read by a native English speaker as training material. Test speech text consisted of 206 words read by the same speaker, and randomly taken from the TIMIT vocabulary with no attempt to maximise the frequency of the words in the training text.

Four different noises were used: random Gaussian noise, aircraft noises (F-16 and Lynx helicopter) from NOISEX database [5], and vehicle noise recorded from inside a moving city bus. Noises were subsequently added to the speech waveform at various SNR: 0, 10, 20, 30 and 40dB. Experiments were performed for each noise type and for each SNR.

Speech signals were sampled at 16kHz and a 13^{th} order MFCC was applied, with a frame shift of 10 ms using a window length of 25.6 ms. For LDA evaluation, ΔMFCC computed using a classical regression were added to the MFCC. We projected this 26-dimensional vector space to a 16-dimensional vector space. No attempt was made to optimise the projection space dimension.

The base transformation between clean reference environment and noisy test environments was computed using nine sentences (20 seconds) from the 130 training sentences, corrupted by noise.

In the STM combination, the noise HMM was trained using 3 seconds of noise.

5.2. Results

The results are in term of % accuracy where for N tokens, S substitution errors, D deletion errors and I insertion errors, accuracy is expressed as $[(N-S-D-I)/N] \times 100\%$. Results for white noise, bus noise, Lynx noise and F16 noise are presented respectively in table 1, 2, 3 and 4. `No comp` means that the training is performed in clean conditions and the test in noisy conditions. Due to our VINICS

Experiment	0 dB	10 dB	20 dB	30 dB	40 dB
No Comp.	–	11.65	48.06	91.26	93.69
Matching	61.17	80.58	89.32	93.69	94.66
STM-1HMM	21.84	73.30	91.26	93.69	95.15
Base Trans.	54.37	86.89	90.29	92.23	92.72
LDA	76.70	88.35	95.15	95.63	96.12

Table 1. Recognition rate vs SNR - Gaussian noise

system strategies, some branches with low probabilities are prematurely cut in the phoneme sequence search. This happens at low SNR and explains why we cannot evaluate recognition performance at 0 dB for the `No comp` experiment. This phenomenon also introduces a slight bias in the performance evaluation at 0 dB for the others experiments. `Matching` denotes identical training and testing environments. This experiment was performed to give the best possible performance that would be obtainable using model compensation schemes. `STM-1HMM` and `STM-2HMM` stand for STM and HMM combination, with respectively a 1 and 2 states HMM. `Base Trans` stands for the base transformation experiment, and `LDA` for experiment with LDA preprocessing, using a matrix transformation computed under the same training and testing conditions.

As expected, the figures show that attempting to recognise noisy speech at low SNR (<10 dB), using a system trained with clean speech leads to a very poor recognition rate, especially for white Gaussian noise.

For high SNRs (30–40 dB), the `STM-HMM` combination outperformed the base transformation method. For such SNRs, it should be noted that there are no significant differences in performance between the `matching` experiment and the `STM-HMM` combination. This confirms that the combination scheme is very effective for a noisy speech recognition task under low level noise. In our experiments, no significant differences were find between `STM-1HMM` and `STM-2HMM`, but this point would probably need further investigations.

For low SNRs (0–10 dB), the performances for the `STM-HMM` experiment drastically fell down, compared to the `matching` experiment. It was recently shown [6] that the noisy speech pdf in the cepstral domain and for low SNR follows a bimodal density distribution. Our log-normal assumption, which led to a normal pdf for the noisy speech in the cepstral domain, is then no longer valid. The compensation scheme assumes that noisy and clean speech models are taken from a similar pdf family, and that they only differ on their respective parameters. From the results in [6], we argue that noisy and clean speech parameters are not taken from a similar pdf family in the cepstral domain, so our compensated noisy speech model does not fit the noisy data, which can explain the bad recognition performance.

The base transformation approach gave very good results at low SNRs, which confirms results obtained in [3]. At 10 dB, the base transformation clearly outperformed the `matching` experiment, for all kind of noises.

The LDA approach generally gave the best results. We recall that the transformation is computed on noisy training data, and applied to the noisy test data. So, LDA recognition performances should be compared with the `Matching` experiment, where the training and testing conditions are identical. It should be interesting to study the robustness of the LDA transformation, trained at a given SNR, and applied to test speech with a different SNR. These experiments are under way.

6. CONCLUSION

Three approaches to noisy speech recognition were compared: a model compensation scheme, a space mapping technique and a preprocessing based on LDA.

The LDA preprocessing showed to be very efficient for the different kinds of noise and SNRs. However, it should be noticed that, in this approach, the training and testing environment are identical. Thus, further experiments have to be carried out to study the robustness of the LDA transformation to changes in the test SNR.

The model compensation scheme performed well under low level noise environment, and gave results similar to those obtained when the training and testing environmental conditions are well matched. At low SNRs, the basic assumption about pdfs combination revealed as being no longer valid, and led to a noisy speech model which does not fit the noisy data. We then obtained poor recognition performance.

The base transformation proved its efficiency at high level noise. At high SNRs, we observed a slight decrease in the recognition rate, in relation to the experiments where the training and testing conditions are identical.

Experiment	0 dB	10 dB	20 dB	30 dB	40 dB
No Comp.	–	68.93	94.66	94.66	94.66
Matching	70.39	83.50	92.72	94.17	96.12
STM-1HMM	32.04	85.44	93.69	95.15	94.66
STM-2HMM	38.83	85.44	93.69	94.17	94.66
Base Trans.	73.79	88.35	91.75	93.69	93.69
LDA	65.53	92.23	94.66	96.12	96.12

Table 2. Recognition rate vs SNR - Bus noise

Experiment	0 dB	10 dB	20 dB	30 dB	40 dB
No Comp.	–	42.23	91.26	93.69	94.66
Matching	76.21	86.89	93.20	93.69	94.17
STM-1HMM	40.29	83.98	92.23	95.63	95.63
STM-2HMM	42.23	85.44	92.72	95.15	95.15
Base Trans.	77.67	88.35	92.72	94.66	93.20
LDA	78.64	93.69	94.17	95.63	98.06

Table 3. Recognition rate vs SNR - Lynx noise

REFERENCES

[1] Y. Gong and J.-P. Haton. Stochastic Trajectory Modeling For Speech Recognition. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 57–60, Adelaide, Australia, April 1994. ICASSP'94.

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

[3] W. C. Treurniet and Y. Gong. Noise Independent Speech Recognition for a Variety of Noise Types. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 437–440, Adelaide, Australia, April 1994. ICASSP'94.

[4] B. S. Atal. Automatic Speaker Recognition Based on Pitch Contours. *Journal of the Acoustical Society of America*, 52(6):1687–1697, 1972.

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

[6] J. Openshaw and J. S. Mason. On the limitations of cepstral features in noise. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 49–52, Adelaide, Australia, 1994. ICASSP'94.

Experiment	0 dB	10 dB	20 dB	30 dB	40 dB
No Comp.	–	29.13	88.83	95.15	95.15
Matching	62.62	83.01	90.29	95.15	94.66
STM-1HMM	13.59	77.18	89.81	94.66	95.15
STM-2HMM	10.68	74.27	89.81	95.15	94.66
Base Trans.	61.65	85.92	92.31	92.72	93.69
LDA	66.02	88.35	93.63	96.12	97.09

Table 4. Recognition rate vs SNR - F16 noise