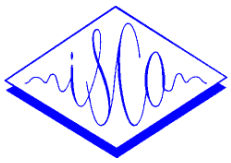

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

AN HMM-BASED CEPSTRAL-DOMAIN SPEECH ENHANCEMENT SYSTEM

C. W. Seymour *M. Niranjan*

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

ABSTRACT

This paper describes a method of enhancing speech corrupted by additive uncorrelated noise. The approach adopted is to use cepstral-domain hidden Markov models to determine statistics of the clean speech and noise processes. A compensated model of speech corrupted by noise is generated using parallel model combination. MMSE and linear nonhomogeneous estimators of the clean speech signal are derived. The enhancement system gives natural-sounding speech without the artifacts introduced by systems such as spectral subtraction. HMM recognition tests performed on the enhanced speech using the NOISEX-92 database show a significant reduction in error rate.

1. INTRODUCTION

This paper is concerned with the problem of enhancing speech corrupted by additive uncorrelated noise. The uses of enhancement are: a) to construct a signal which, according to some criteria, sounds better than the noisy speech; and b) to estimate some parameters of the clean speech as a front end to a computer speech recognition system. When an enhanced speech signal is to be constructed, the resulting speech can be judged according to criteria such as intelligibility, quality and the suppression of noise. Both forms of enhancement are considered here and results are presented using the enhancement system as the front end for an HMM-based speech recogniser.

The assumptions about the nature of the speech and noise processes depend on the enhancement method being used. For example, the iterative Wiener filtering scheme assumes that the speech is produced by an autoregressive process [5]. Spectral subtraction however, only explicitly assumes the nature of the noise power spectrum [1]. HMM-based enhancement systems make the additional assumption that the speech is formed by a Markov process. HMM-based systems have been developed based on an autoregressive model of speech [2] and a filterbank model [3].

State of the art recognisers have been shown to perform well using a model of speech in the cepstral domain. In this work, a cepstral model is used as this can be expected to give improved frame-to-state mapping and hence an improved estimate of the clean speech statistics. A fully-connected, cepstral-domain HMM is trained on clean speech and a noise model is trained on noise in the absence of speech. These are combined to form a model of noise-corrupted speech using cepstral parameter compensation and parallel model combination [4]. To perform enhancement, a Baum-Welch pass is first performed on the noisy speech using the compensated model, yielding the state/mixture occupancy probabilities at each time frame. These are used in conjunction with the statistics from the clean speech and noise models to form an estimate of the clean-speech short-time power spectra.

MMSE and linear nonhomogeneous estimators of the speech short-time power spectrum are derived. Vocabulary-dependent, speaker-dependant and vocabulary-independent, speaker-independent speech models are compared.

2. SIGNAL PARAMETRISATION

Let $\mathbf{s}_t^{\mathrm{t}} \in \mathbb{R}^N$ be the clean speech vector at time frame t, multiplied by a suitable window function. Similarly, let $\mathbf{n}_t^{\mathrm{t}}$ represent the noise signal and $\mathbf{y}_t^{\mathrm{t}}$ represent the noisy speech signal. The noise is assumed to be additive, therefore

$$\mathbf{y}_t^{\mathrm{t}} = \mathbf{s}_t^{\mathrm{t}} + \mathbf{n}_t^{\mathrm{t}} \tag{1}$$

In this work, a superscript t indicates the time domain, l indicates the log-spectral domain and c indicates the cepstral domain. The representation in the short-time power-spectral domain is given by

$$\mathbf{s}_t = |\mathbf{F}\mathbf{s}_t^{\mathrm{t}}|^2 \tag{2}$$

where $\mathbf{s}_t \in \mathbb{R}^M$, $M = N/2+1$ and $\mathbf{F} \in \mathbb{R}^{M\times N}$ is the discrete Fourier transform matrix. The relationship between the signals in this domain is given by

$$y_i = s_i + n_i + 2\sqrt{s_i n_i}\cos\theta_i \tag{3}$$

where s_i is an element of $\mathbf{s}_t$ and θ_i is the angle between elements of $\mathbf{F}\mathbf{s}_t^{\mathrm{t}}$ and $\mathbf{F}\mathbf{n}_t^{\mathrm{t}}$. The cross term will be ignored in this work; hence the approximation

$$\mathbf{y}_t = \mathbf{s}_t + \mathbf{n}_t \tag{4}$$

The signals are modelled by HMMs in the cepstral domain,

$$\mathbf{s}_t^{\mathrm{c}} = \mathbf{C}\mathbf{s}^{\mathrm{l}} \tag{5}$$

where $\mathbf{C} \in \mathbb{R}^{M\times M}$ is given by

$$C_{ij} = \frac{1}{N}\begin{cases} 1 & \text{if } j = 1 \\ (-1)^{i-1} & \text{if } j = M \\ 2\cos(2(i-1)(j-1)\pi/N) & \text{otherwise} \end{cases} \tag{6}$$

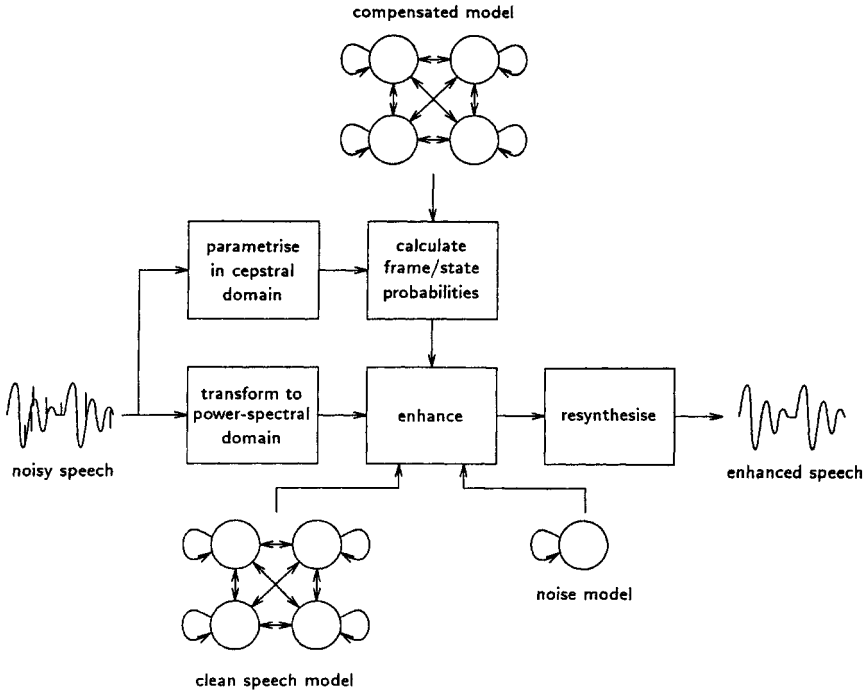

Figure 1. Enhancement system

3. HMM NOISE COMPENSATION

Given models trained on the clean speech and noise signals, a model for the noisy speech is required. This consists of a set of composite states, representing the noisy speech when the clean speech and noise models are in a given state. An accent $\bar{}$ indicates parameters of the noise process, an accent $\tilde{}$ indicates the noisy speech process and the absence of an accent indicates the clean speech process. The transition probabilities for the noise-compensated model are given by

$$\bar{a}_{[iu][jv]} = a_{ij}\tilde{a}_{uv} \tag{7}$$

The state $[iu]$ of the of the compensated model represents the noisy speech when the clean speech model is in state i and the noise model is in state u.

A method for determining the distributions of the compensated model states has been devised by Gales [4]. This uses the assumption that the speech and noise signals are additive in the short-time power-spectral domain. Let $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ be the the mean and covariance of an HMM state.

Given that the distributions in the cepstral domain are Gaussian, the distributions in the log-spectral domain are Gaussian with means and variances given by

$$\boldsymbol{\mu}^{\mathrm{l}} = \mathbf{C}^{-1}\boldsymbol{\mu}^{\mathrm{c}} \tag{8}$$

$$\boldsymbol{\Sigma}^{\mathrm{l}} = \mathbf{C}^{-1}\boldsymbol{\Sigma}^{\mathrm{c}}\mathbf{C}^{-1^{\mathrm{T}}} \tag{9}$$

The distributions in the power-spectral domain have parameters

$$\mu_i = \exp(\mu_i^{\mathrm{l}} + \Sigma_{ii}^{\mathrm{l}}/2) \tag{10}$$

$$\Sigma_{ij} = \mu_i\mu_j[\exp(\Sigma_{ij}^{\mathrm{l}}) - 1] \tag{11}$$

Since the signals are assumed to be additive in the power-spectral domain (equation 4), the mean and covariance of the compensated model are given by

$$\bar{\boldsymbol{\mu}} = \boldsymbol{\mu} + \tilde{\boldsymbol{\mu}} \tag{12}$$

$$\bar{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} + \tilde{\boldsymbol{\Sigma}} \tag{13}$$

If the distribution of the noisy signal in the power-spectral domain is assumed to be lognormal, then the parameters in the cepstral domain can be found by inverting the above transforms.

$$\bar{\Sigma}_{ij}^{\mathrm{l}} = \ln\left[\frac{\bar{\Sigma}_{ij}}{\bar{\mu}_i\bar{\mu}_j} + 1\right] \tag{14}$$

$$\bar{\mu}_i^{\mathrm{l}} = \ln(\bar{\mu}_i) - \frac{1}{2}\bar{\Sigma}_{ii}^{\mathrm{l}} \tag{15}$$

and

$$\bar{\boldsymbol{\mu}}^{\mathrm{c}} = \mathbf{C}\bar{\boldsymbol{\mu}}^{\mathrm{l}} \tag{16}$$

$$\bar{\boldsymbol{\Sigma}}^{\mathrm{c}} = \mathbf{C}\bar{\boldsymbol{\Sigma}}^{\mathrm{l}}\mathbf{C}^{\mathrm{T}} \tag{17}$$

A better estimate of the compensated model parameters can be made without making this approximation. An element of the observed speech vector in the log-spectral domain is given by

$$y_i^{\mathrm{l}} = \ln(\exp(s_i^{\mathrm{l}}) + \exp(n_i^{\mathrm{l}})) \tag{18}$$

The components of $\bar{\boldsymbol{\mu}}^{\mathrm{l}}$ and $\bar{\boldsymbol{\Sigma}}^{\mathrm{l}}$ are given by the mean of y_i^{l} and covariance of y_i^{l} and y_j^{l} respectively. Therefore, $\bar{\boldsymbol{\mu}}^{\mathrm{l}}$ is given by

$$\bar{\mu}_i^{\mathrm{l}} = \int\int_{-\infty}^{\infty} \ln(\exp(s_i^{\mathrm{l}}) + \exp(n_i^{\mathrm{l}}))p(s_i^{\mathrm{l}})p(n_i^{\mathrm{l}})\,ds_i^{\mathrm{l}}\,dn_i^{\mathrm{l}} \tag{19}$$

where $p(s_i^{\mathrm{l}})$ is Gaussian with mean and variance given by equations 8 and 9. Similarly, an expression can be written for $\bar{\boldsymbol{\Sigma}}^{\mathrm{l}}$. The cepstral-domain parameters $\bar{\boldsymbol{\mu}}^{\mathrm{c}}$ and $\bar{\boldsymbol{\Sigma}}^{\mathrm{c}}$ can be obtained using equations 16 and 17.

Diagonal covariance matrices were used in the clean speech and noise models. In the power-spectral and cepstral domains, these are not diagonal. Compensation generates full covariance matrices, however

the off-diagonal elements in the compensated model were neglected. There is no closed form solution to equation 19, therefore it must be evaluated by numerical integration. This method was used to calculate the compensated mean vectors but the lognormal approximation was used to calculate the compensated covariance matrices. This was due to the computational cost of estimating each element of $\bar{\mathbf{\Sigma}}^{l}$, since the off-diagonal elements can only be neglected in $\bar{\mathbf{\Sigma}}^{c}$.

4. SIGNAL ESTIMATION

The enhancement system is shown in figure 1. The noisy speech is parametrised in the cepstral domain and the probability that the compensated HMM is in a given state at each time frame is calculated using a Baum-Welch pass. Each state of the compensated HMM corresponds to a pair of states from the clean speech and noise models. The means and covariances of these states are known, therefore for a given pair of states, an estimate of $\mathbf{s}_t$ can be derived. The speech estimator is formed by combining these estimators with a weighting given by the probability that the HMM is in each given state.

The enhanced power spectra are combined with the phase of the noisy speech spectra and transformed with an inverse DFT. This yields frames of estimated speech in the time domain which can be overlapped and added to reconstruct the speech signal. For the recognition task, the enhanced power spectra are converted into MFCC parameters which are input to an HMM speech recogniser trained on clean speech.

The posterior clean speech pdf, given the noisy speech, is

$$p(\mathbf{s}_t|\mathbf{Y}) = \sum_i P(Q_{ti}|\mathbf{Y}) p(\mathbf{s}_t|Q_{ti}, \mathbf{y}_t) \tag{20}$$

where Q_{ti} represents the event that the compensated model is in state i at time t. The posterior state probabilities $P(Q_{ti}|\mathbf{Y})$ may be obtained from the forward and backward probabilities.

$$P(Q_{ti}|\mathbf{Y}) = \frac{p(\mathbf{Y}, Q_{ti})}{p(\mathbf{Y})} \tag{21}$$

$$= \frac{\alpha_i(t)\beta_i(t)}{\sum_i \alpha_i(t)\beta_i(t)} \tag{22}$$

These probability densities are given by

$$\alpha_i(t) = p(\mathbf{y}_{1,t}, Q_{ti}) \tag{23}$$

$$\beta_i(t) = p(\mathbf{y}_{t+1,T}|Q_{ti}) \tag{24}$$

The posterior clean speech pdf, given the HMM state, is given by

$$p_s(\mathbf{s}_t|Q_{ti}, \mathbf{y}_t) \tag{25}$$

$$= \frac{p_s(\mathbf{s}_t|Q_{ti}) p_y(\mathbf{y}_t|\mathbf{s}_t, Q_{ti})}{p_y(\mathbf{y}_t|Q_{ti})} \tag{26}$$

$$= \frac{p_s(\mathbf{s}_t|Q_{ti}) p_n(\mathbf{y}_t - \mathbf{s}_t|Q_{ti})}{p_y(\mathbf{y}_t|Q_{ti})} \tag{27}$$

For the multiple-mixture case, the posterior state mixture probability $P(Q_{ti}, M_{tj}|\mathbf{Y})$ is required, where M_{tj} represents the event that the model is in mixture component j at time t.

$$P(Q_{ti}, M_{tj}|\mathbf{Y}) \tag{28}$$
$$= P(Q_{ti}|\mathbf{Y}) P(M_{tj}|Q_{ti}, \mathbf{y}_t)$$

$$P(M_{tj}|Q_{ti}, \mathbf{y}_t) \tag{29}$$
$$= \frac{p(\mathbf{y}_t|Q_{ti}, M_{tj}) P(M_{tj}|Q_{ti})}{\sum_j p(\mathbf{y}_t|Q_{ti}, M_{tj}) P(M_{tj}|Q_{ti})}$$

For the remainder of this section, single mixture models will be assumed. The extension to the multiple mixture case is trivial.

4.1. MMSE Estimation

The MMSE estimator of $\mathbf{s}_t$ is given by

$$\hat{\mathbf{s}}_t = E\{\mathbf{s}_t|\mathbf{Y}\} \tag{30}$$

$$= \int \mathbf{s}_t p(\mathbf{s}_t|\mathbf{Y})\, d\mathbf{s}_t \tag{31}$$

$$= \sum_i P(Q_{ti}|\mathbf{Y}) \int \mathbf{s}_t p(\mathbf{s}_t|Q_{ti}, \mathbf{y}_t)\, d\mathbf{s}_t$$

where $P(Q_{ti}|\mathbf{Y})$ is given by equation 22 and $p(\mathbf{s}_t|Q_{ti}, \mathbf{y}_t)$ is given by equation 27. The pdf $p(\mathbf{s}_t|Q_{ti})$ can be derived from $p^{l}(\mathbf{s}_t|Q_{ti})$ which is Gaussian with mean μ^{l} and covariance Σ^{l} given by equations 8 and 9. If the off-diagonal terms of Σ^{l} are neglected then the pdf of an element of $p(\mathbf{s}_t|Q_{ti})$ is

$$p(s|Q_{ti}) = \frac{1}{\sigma^{l} s\sqrt{2\pi}} \exp - \frac{(\ln s - \mu^{l})^2}{2\sigma^{l\,2}} \tag{32}$$

where $\sigma^{l\,2} = \Sigma^{l}_{ii}$. Therefore, from equation 27,

$$p(s|Q_{ti}, y) = \frac{1}{2\pi p(y|Q_{ti}) s^2 \sigma^{l} \tilde{\sigma}^{l}} \tag{33}$$
$$\exp - \left(\frac{(\ln s - \mu^{l})^2}{2\sigma^{l\,2}} + \frac{(\ln(y-s) - \tilde{\mu}^{l})^2}{2\tilde{\sigma}^{l\,2}} \right)$$

and

$$p(y|Q_{ti}) = \frac{1}{2\pi\sigma^{l}\tilde{\sigma}^{l}} \tag{34}$$
$$\int_0^y \frac{1}{s^2} \exp - \left(\frac{(\ln s - \mu^{l})^2}{2\sigma^{l\,2}} + \frac{(\ln(y-s) - \tilde{\mu}^{l})^2}{2\tilde{\sigma}^{l\,2}} \right) ds$$

This estimator has the disadvantage that there is no closed form solution. It may be approximated by means of numerical integration, however, this will be computationally intensive.

4.2. Linear Nonhomogeneous Estimator

A linear form of the speech estimator is given by

$$\hat{\mathbf{s}}_t = \mathbf{A}\mathbf{y}_t + \mathbf{b} \tag{35}$$

To simplify the problem, the off-diagonal terms of the covariance matrix $\mathbf{\Sigma}$ are again neglected, so the estimator of an element of $\mathbf{s}_t$ takes the form

$$\hat{s} = Ay + B \tag{36}$$

The estimator is taken to be the weighted sum of estimators given each state of the speech model.

$$\hat{s} = \sum_i P(Q_{ti}|\mathbf{Y})(A_i y + B_i) \tag{37}$$

For a given state i, the mean square error is minimised if

$$A_i = r_i \frac{\sigma_i}{\bar{\sigma}_i} \quad (38)$$

$$B_i = \mu_i - A_i \bar{\mu}_i \quad (39)$$

where

$$r_i = \frac{E\{sy\} - E\{s\}E\{y\}}{\sigma_i \bar{\sigma}_i} \quad (40)$$

Hence,

$$A_i = \frac{\sigma_i^2}{\sigma_i^2 + \tilde{\sigma}_i^2} \quad (41)$$

$$B_i = \mu_i - A(\mu_i + \tilde{\mu}_i) \quad (42)$$

This estimator is used by Gagnon [3] and has the advantage of being less computationally intensive.

5. SPEECH MODEL

The speech models were trained using Baum-Welch re-estimation. Multiple-mixture models were used, and training was performed by gradually increasing the number of mixture components for each state and re-estimating the model at each stage. Two types of speech model are considered here. A vocabulary-dependent, speaker-dependent model was trained on digits spoken by a single speaker. The same speaker was used for the recognition task. The vocabulary-independent, speaker-independent model was trained on a total of 25 seconds of speech spoken by 4 male speakers from the WSJCAM0 database.

A single state of the speech model was used to represent silence. The compensated form of this state is the same as the noise model. For the recognition task, frames for which the probability of the model being in the silence state was greater than a given threshold were labelled as silence. During the Viterbi recognition stage, the state output density for frames classified as silence was set to a constant for speech model states and to a larger constant for silence model states.

6. RESULTS

In this section, recognition results are reported on the NOISEX-92 database, using the enhancement system used as the front end to an HMM-based speech recogniser. A 25 msec frame size and 10 msec frame period were used. The parametrisation used by the recogniser was a set of 15 MFCC coefficients. Diagonal covariance matrices were used by both the enhancement model and the recogniser models. The enhancement model consisted of 8 speech states with 7 mixture components each and a silence state with a single mixture component. Single state, single mixture, noise models were used.

Both the MMSE and the linear nonhomogeneous estimators were implemented. The enhanced speech from the two estimators sounded very similar, however the recognition results using the linear nonhomogeneous estimator were superior. Therefore, the more complex MMSE estimator was rejected.

Recognition results in terms of percentage accuracy for the NOISEX-92 isolated digit recognition task are shown in table 1. The vocabulary dependent speaker dependent (VDSD) enhancement system can be seen to perform better than the vocabulary independent speaker independent (VISI) system.

SNR /dB		-6	0	6	12	18
Car	Baseline	15	30	59	73	82
	VDSD	56	89	99	100	100
	VISI	26	63	87	98	100
Lynx	Baseline	10	17	49	72	98
	VDSD	64	94	100	100	100
	VISI	34	66	93	100	100

Table 1. Results for NOISEX-92 isolated digit task

In informal listening tests, the enhancement system proved to be very effective at removing the interfering noise, with very little residual noise remaining. At signal to noise ratios down to 6 dB, the quality of the enhanced speech was good. Under noisier conditions, the enhanced speech became rather coarse, with lower energy regions of the speech signal being over attenuated. No artifacts such as the "musical noise" associated with spectral subtraction were present in the enhanced speech.

7. CONCLUSION

An approach was developed for enhancing noisy speech based on cepstral-domain HMM modelling of the speech and noise processes. A simple linear non-homogenous estimator of the speech short-time power spectrum was found to perform well. The enhancement scheme was successful both as a preprocessor for an HMM-based speech recogniser and for generating an enhanced speech waveform. In HMM recognition tests, the error rate was substantially reduced.

REFERENCES

[1] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-27:113–120, 1979.

[2] Y. Ephraim. A Bayesian estimation approach for speech enhancement using hidden Markov models. *IEEE Trans. Signal Processing*, 40(4):725–735, 1992.

[3] L. Gagnon. A noise reduction approach for non-stationary additive interference. *Proc. ETRW*, pages 139–142, 1992.

[4] M. J. F. Gales and S. J. Young. Cepstral parameter compensation for HMM recognition in noise. *Speech Communication*, 12:231–240, 1993.

[5] J. S. Lim and A. V. Oppenheim. All-pole modeling of degraded speech. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-26:197–210, 1978.