Combining a Gaussian Mixture Model Front End with MFCC Parameters

Matthew Stuttle & Mark Gales

19th September 2002

Cambridge University Engineering Department

Web: http://svr-www.eng.cam.ac.uk/~mns25
Email: mns25@eng.cam.ac.uk

Overview

- The GMM speech frontend
- Confidence metrics
- Compensation of GMM parameters in a noisy enviroment
- Experimental results
  - Confidence Metric
  - Noise corrupted systems
  - Noise compensation approaches
- Conclusions

The Gaussian Mixture Model in feature extraction

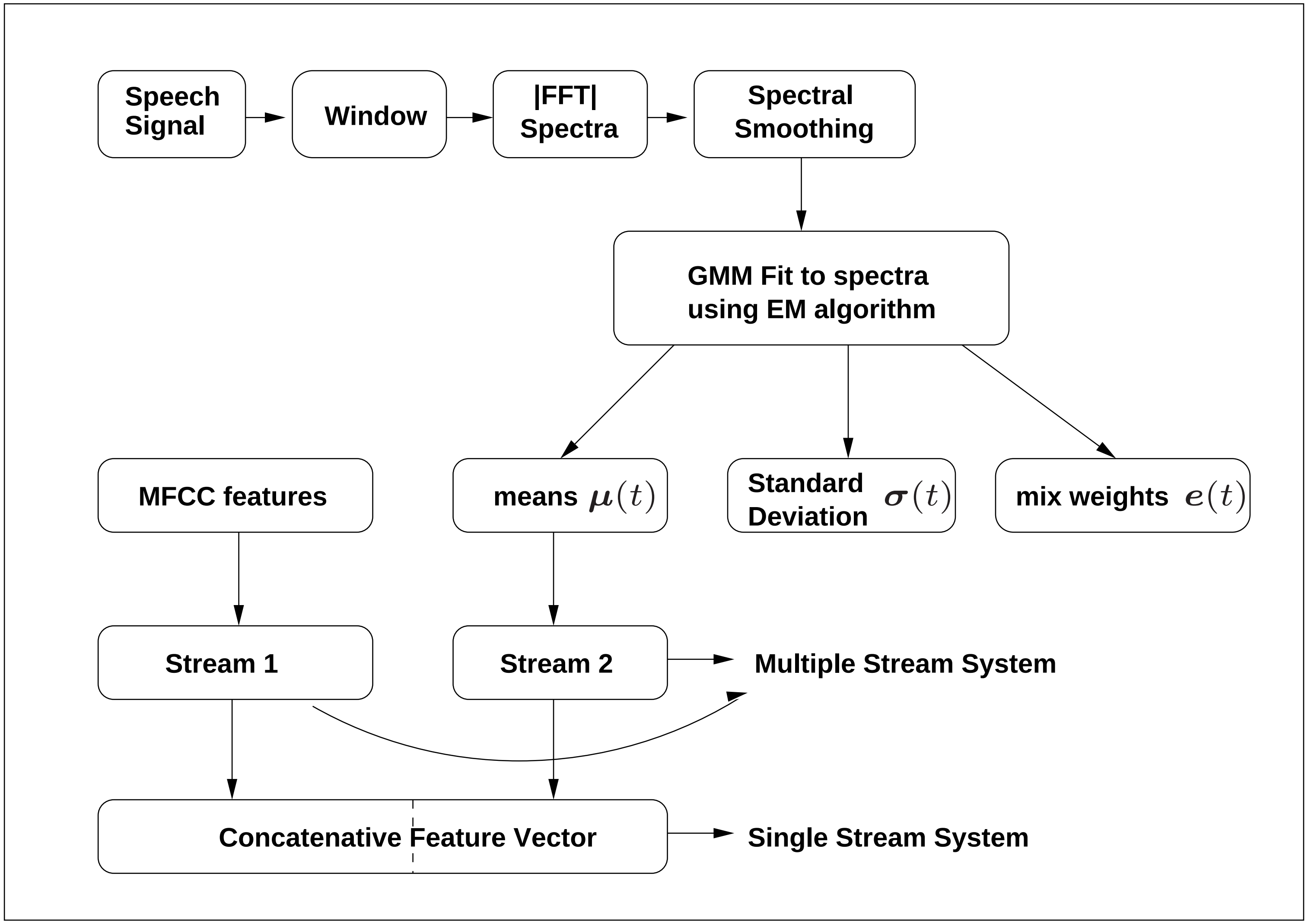
Problems with existing formant extraction schemes:

- Not always well defined in spectra, (eg fricatives or nasalised sounds)
- Amplitude information required to distinguish certain phone types (eg nasalised phones and voiced vowels)

Gaussian mixture model:

- Fits a set of Gaussian mixtures to the smoothed magnitude spectrum of a speech signal
- Characterises the spectrum in terms of spectral peaks, hence the features are ‘formant-like’

Gaussian Mixture Model front end



Confidence in GMM Fit Metrics

- Peaks are less reliably defined in unvoiced or quiet regions
- Define confidence metric $\xi(t)$ based on amplitude and curvature

$$\xi(t) = \beta \left[\prod_{n=1}^{N} \frac{e_n(t) + 10.53}{\sigma_n(t)} \right]^{\frac{1}{N}}$$

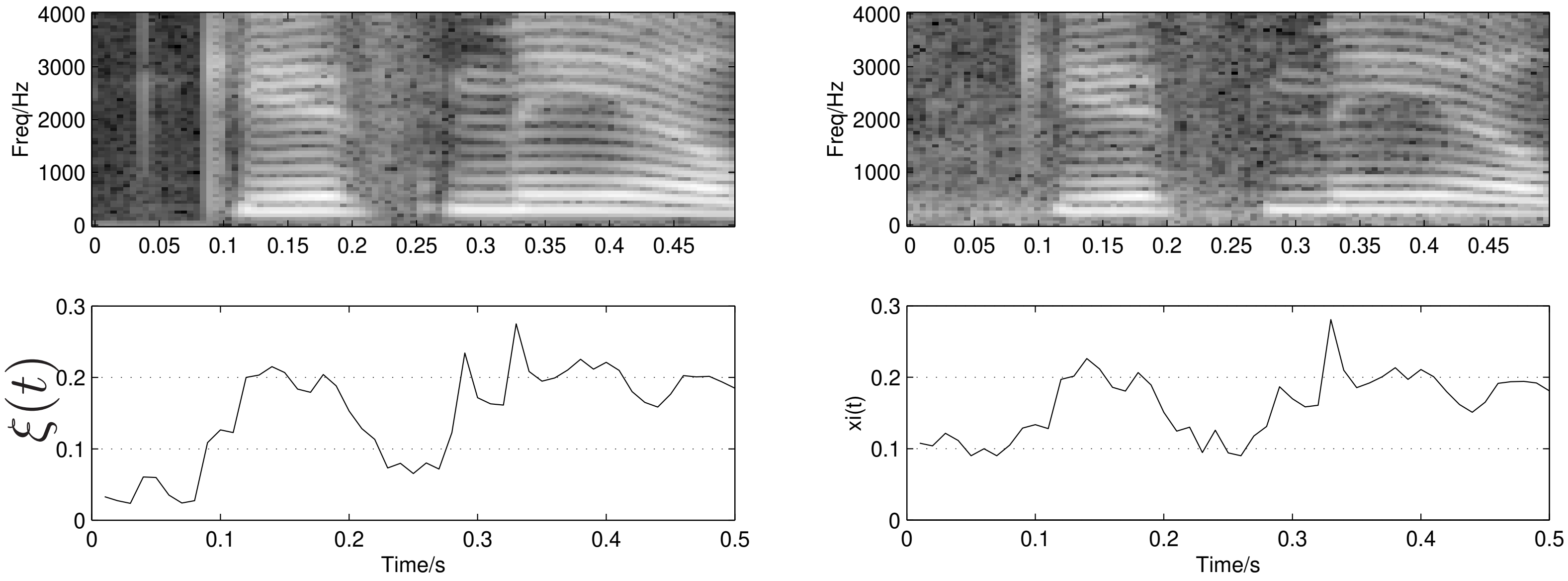
- Use standard synchronous stream system

$$b_j(\boldsymbol{y}(t)) = \prod_{r=1}^{R} \left[\sum_{m=1}^{M} c_{jrm} \mathcal{N}(\boldsymbol{y}_r(t); \boldsymbol{\mu}_{jrm}, \boldsymbol{\Sigma}_{jrm}) \right]^{\gamma_r(t)}$$

- Stream weights $\gamma_r(t)$ set by confidence metric

$$\gamma_1(t) = 1 - \xi(t) \qquad \gamma_2(t) \propto \xi(t))$$

Example Confidence Metric



- Clean and noise-corrupted plots shown
- $\xi(t)$ is high in regions with peak-structures
- Is low in regions with low energy or no peaks

Front End Noise Compensation

- Compensate at feature extraction stage
- Assumes noise model $\hat{\boldsymbol{\theta}}^{(n)} = \{\hat{\boldsymbol{e}}^{(n)}, \hat{\boldsymbol{\mu}}^{(n)}, \hat{\boldsymbol{\sigma}}^{(n)}\}$
- Estimate clean speech feature parameters given noise model

$$l(\boldsymbol{x}(t)|\boldsymbol{\theta}(t), \hat{\boldsymbol{\theta}}^{(n)}) =$$

$$\sum_{k=1}^{K} \ln \left(\sum_{q=1}^{Q} \hat{e}_q^{(n)} \mathcal{N}\left(x_k(t); \hat{\mu}_q^{(n)}, \hat{\sigma}_q^{(n)2}\right) + \sum_{n=1}^{N} e_n(t) \mathcal{N}\left(x_k(t); \mu_n(t), \sigma_n^2(t)\right) \right)$$

Model Compensation

- Adapts the static mean parameters of clean HMM model trained on GMM parameters
- Reconstructs spectra $\boldsymbol{x}_{jm}$ from GMM parameters of each state j and component m in model
- Noise corrupted spectra is formed by adding spectra from noise spectrum $\boldsymbol{q}$
- Parameters for noisy data $\hat{\boldsymbol{\theta}}_{jm}$ are re-estimated

$$l(\boldsymbol{x}_{jm} + \boldsymbol{q} | \hat{\boldsymbol{\theta}}_{jm}) = \sum_{k=1}^{K} \left(\ln \sum_{n=1}^{N} \hat{e}_{jmn} \mathcal{N} \left(x_{jmk} + q_k ; \hat{\mu}_{jmn}, \hat{\sigma}^2_{jmn} \right) \right)$$

Experimental setup

WSJ task

- 284 training speakers, 65,000 word vocabulary, Hub 1 dev and eval
- Cross-word triphone context-dependent HMMs
- Trigram language model
- Cepstal Mean Normalisation used on feature vectors

RM task

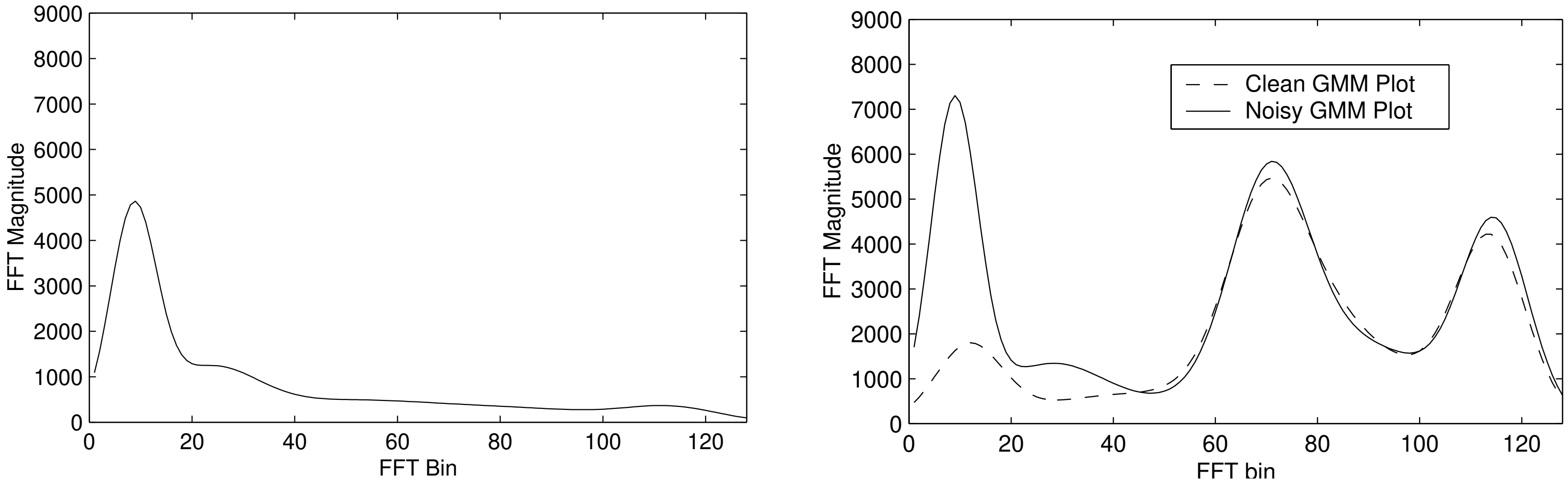
- 3990 training sentences with roughly a 1000 word vocabulary, 109 training speakers and 1200 test sentences from 40 subjects, noise added
- Cross-word triphone context-dependent HMMs used
- A word-pair grammar was used for recognition
- Word Error Rate averages over all 4 test sets quoted

Results on WSJ using confidence metric

Description	% WER
MFCC	9.75
MFCC+6 Means Concatenative	9.56
MFCC+6 Means Fixed Stream Weights	9.64
MFCC+6 Means Confidence Metric	9.52

- Small improvements over fixed stream weights
- No significant improvement over concatenative feature vectors by using confidence metrics on clean speech

Additive Noise



- Noise source is Operations Room noise from the Noisex database
- Data corrupted by adding noise at waveform level
- Coloured noise distrupts peak structure severely
- Noise spectrum and corrupted spectrum shown

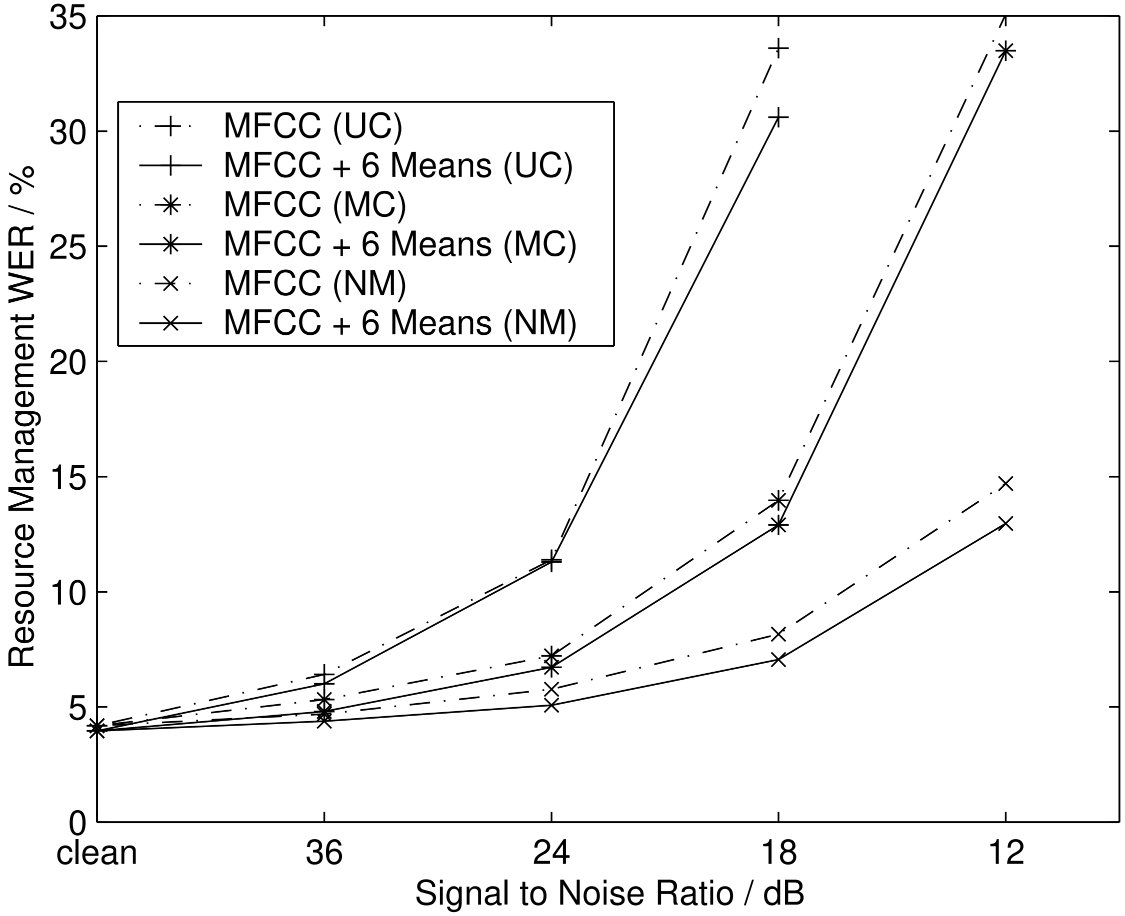
RM Results in additive noise - I

Results using **UC** Uncompensated clean speech models
MC Mean compensated models
NM Noise matched models

18 dB SNR	UC	MC	NM
MFCC	32.3	14.0	8.1
MFCC+GMM Concat.	30.6	13.1	7.1
+ Confidence	29.6	12.6	7.1

- Adding GMM parameters to MFCCs gives improvements in noisy conditions
- Confidence metric yields small improvements for model compensated data
- Frontend compensation to the GMM parameters gave 28.3% WER

RM Results in additive noise - II



- Adding GMM features to MFCCs gives small improvements over a range of SNRs.

Conclusions

- Combining MFCCs with GMM features using synchronous streams measure of confidence yields no significant improvement over concatenating into a single feature vector
- Including GMM features with MFCCs gives improved performance in an additive noise environment
- The static mean parameters of GMM features can be rapidly adapted to additive noise environments

Future Work

- Repeat experiments on speech recorded in noisy enviroments rather than artificially adding noise
- Investigate schemes for adapting Δ and $\Delta\Delta$ parameters.