A Mixture of Gaussians Front End for Speech Recognition

Matthew Stuttle

28th January 2003

Cambridge University Engineering Department

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

Overview

- The GMM speech frontend
  - Motivation
  - Implementation
- Performance of GMM features
  - Baseline results
  - Concatenated with MFCCs
  - Streaming systems
- Confidence metrics
- Noise compensation
- Speaker Adaptation
- Conclusions

The case for formants in LVCSR

Motivation for using formants:

- Considered representative of underlying phonetic content
- Potentially useful in noisy or band-limited enviroments
- Formant positions important for human speech recognition

Existing formant schemes:

- Analysis by synthesis
- Linear prediction analysis
- Dynamic template matching of hand-labelled spectra

Problems with formants

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)

Statistical peak representations:

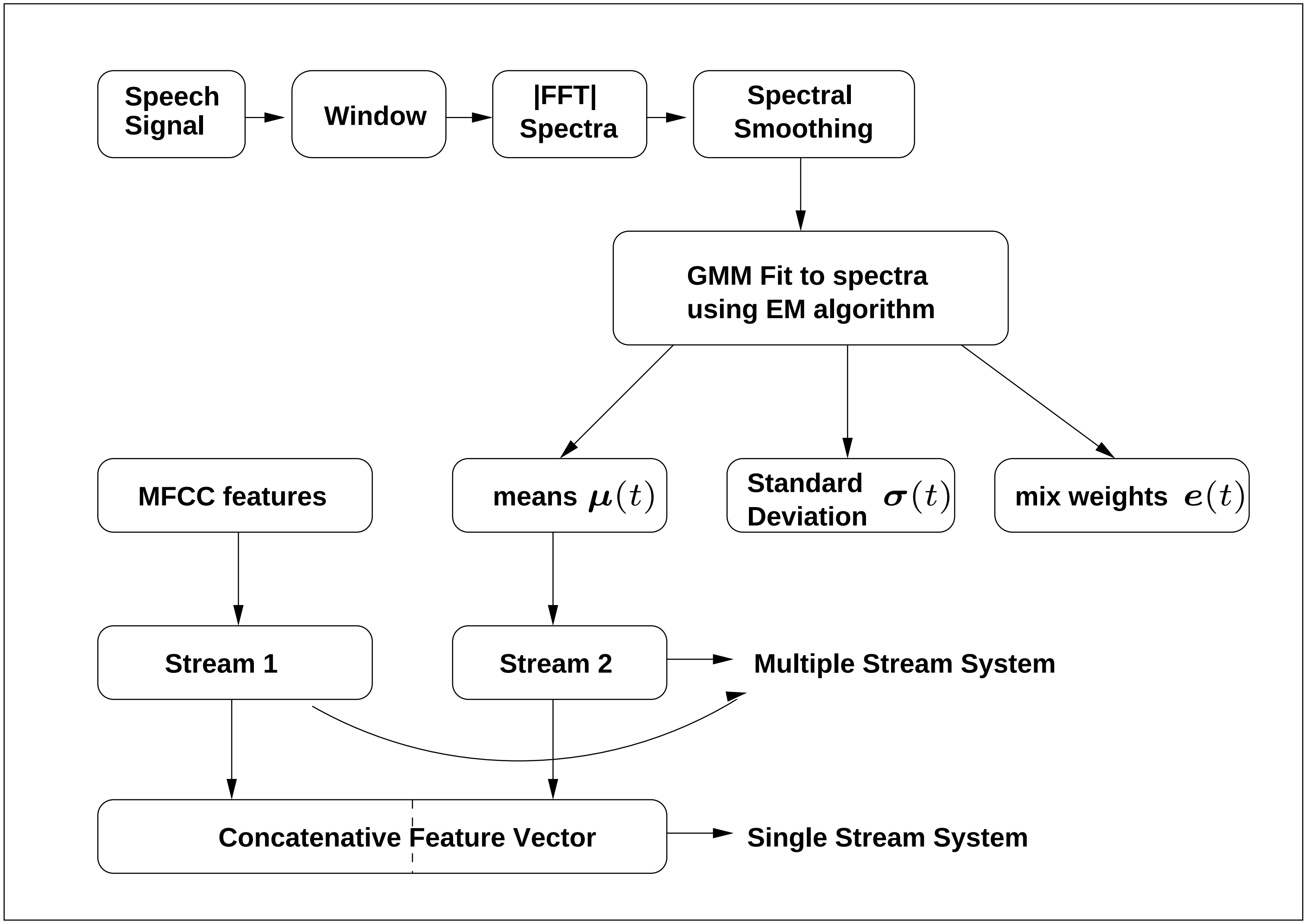
- Gravity Centroids: extract first and second moments from spectral subbands
- HMM-2: fit a second frequency HMM to the spectrum at each frame, each frequency state corresponds to a spectral peak or region

The Gaussian Mixture Model for feature extraction

Gaussian mixture model:

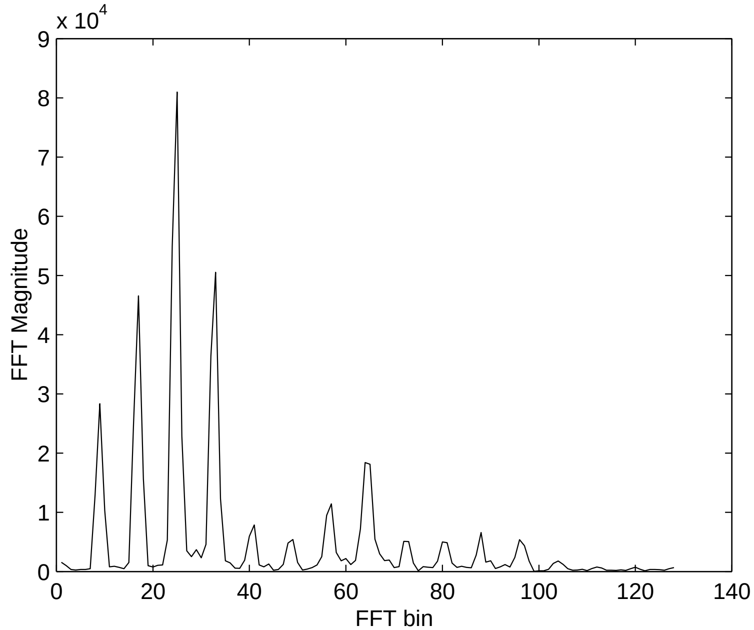
- Fits a set of Gaussian mixtures to the smoothed magnitude spectra of a speech signal
- Characterises the spectra in terms of spectral peaks, hence the features are ‘formant-like’.
- Can represent general spectral envelope
- Statistical representation
- Is not band-limited as Gravity Centriods

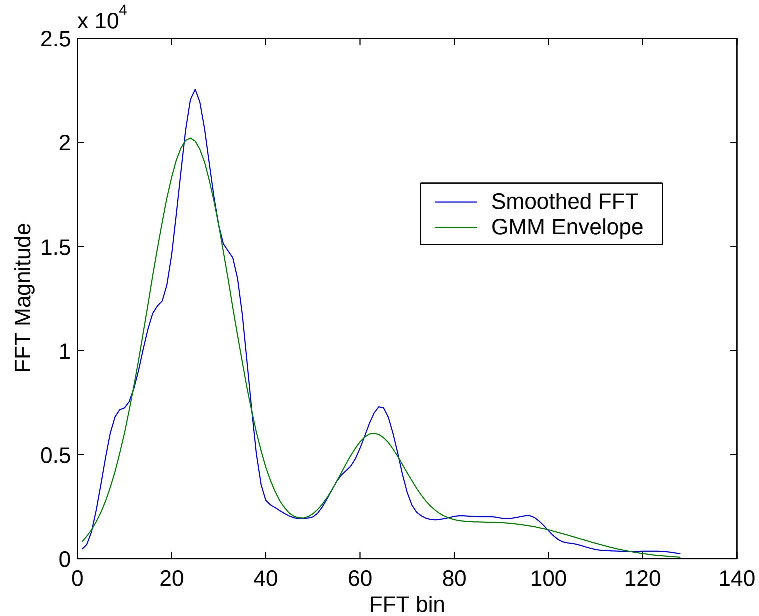
Gaussian Mixture Model front end



Single coded frame

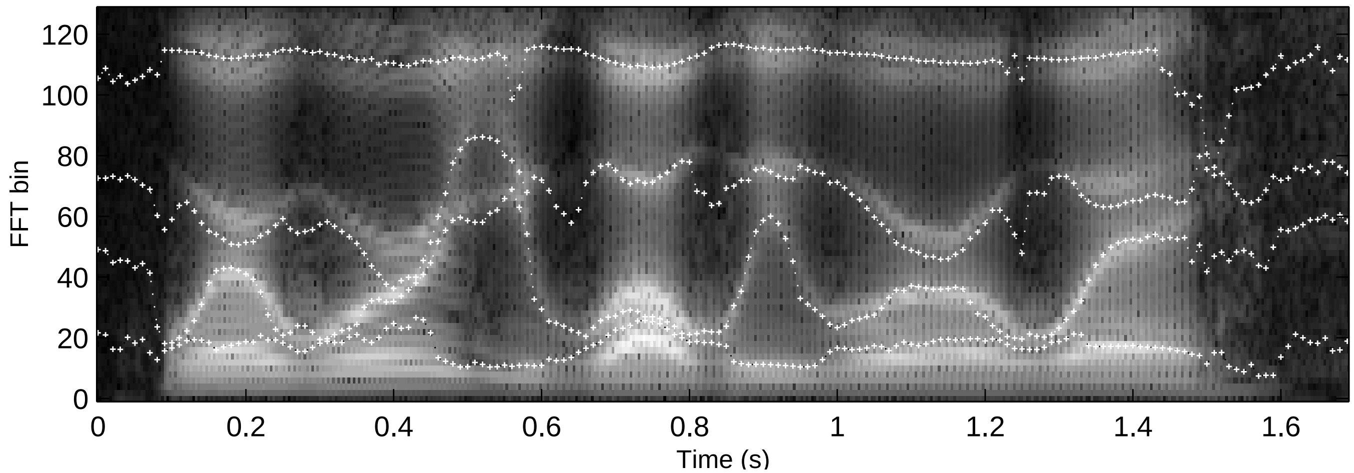
Example single frame plot from test utterance, before and after smoothing.





GMM front end trajectory plot

- Utterance "Where were you while we were away?"
- Four Gaussian components fitted per frame



- Extracts close approximation to formant positions
- No spectral smoothing or frame to frame constraints

Experimental details

All experiments were performed on the Resource Management (RM) task

- 3990 training sentences with roughly a 1000 word vocabulary, 109 training speakers and 1200 test sentences from 40 subjects
- Cross-word triphone context-dependent HMMs were made using a phonetic decision class tree as per HTK RM Recipe
- A word-pair grammar was used for recognition
- Results were tuned on the 300 sentence ‘feb89’ subset of data
- Word Error Rate averages over all 4 test sets quoted

Baseline Resource Management results

Description	Total Features	% WER
MFCC	39	4.19
PLP	39	3.89
4 Component GMM	39	6.10
6 Component GMM	57	4.90

- Best GMM features result was 17% worse than the MFCC baseline
- Fitting six mixtures (GMM6) to spectra yields better result than four
- Errors were distributed evenly across phone classes

Resource Management results for hybrid systems

Gaussian means were appended directly onto the MFCC feature vector

Parameterisation	Total Features	% Err
MFCC $\{c_1 \cdots c_{12}\}$	39	4.19
MFCC + $\{c_1 \cdots c_{16}\}$	51	4.29
MFCC + 4 Formant frequencies from ESPS	51	4.89
MFCC + 4 Gravity Centroids	51	4.08
MFCC + 6 Gravity Centroids	57	5.02
MFCC + 4 GMM Means	51	4.08
MFCC + 6 GMM Means	57	3.96

- Appending the GMM means gave a WER decrease of 5.5% relative to MFCC baseline
- Adding four Gravity Centroids reduced the WER by 2%
- All other features appended degraded performance

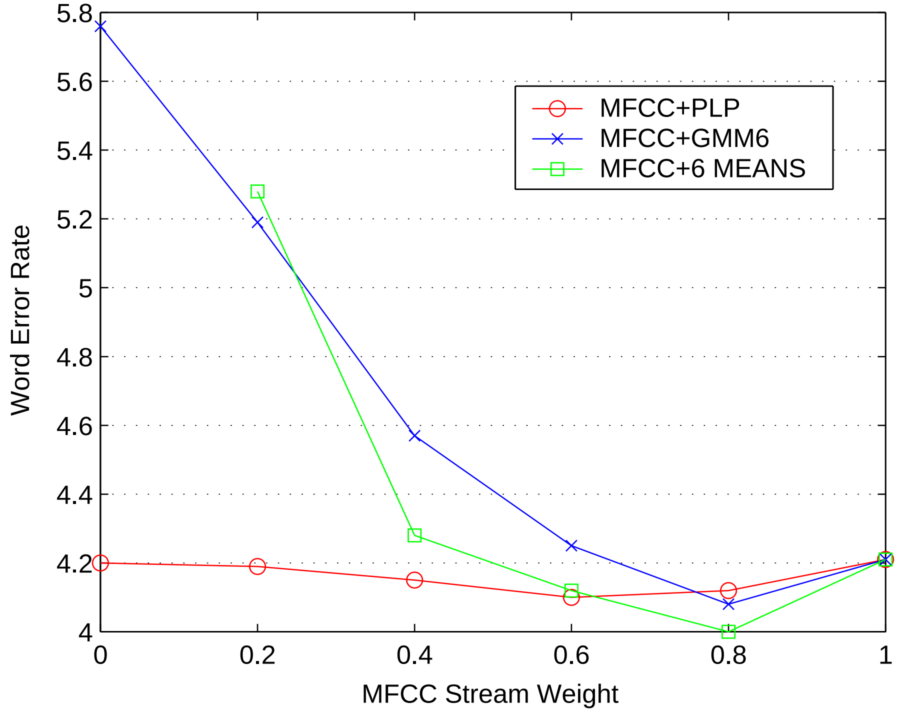
Synchronous stream system

- Input vector $\boldsymbol{y}$ divided into 2 streams $\{\boldsymbol{y}_{MFCC}, \boldsymbol{y}_{GMM}\}$
- Output probability given by

$$b_j(\boldsymbol{y}) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M} c_{jsm} \mathcal{N}(\boldsymbol{y}_s; \boldsymbol{\mu}_{jsm}, \boldsymbol{\Sigma}_{jsm}) \right]^{\gamma_s}$$

- Where $\boldsymbol{\gamma}_s$ is the stream weight of stream s.
- Stream weights were constrained to sum to one.
- Only MFCCs were used to obtain alignments in Baum Welch training

Results for streamed system



- Optimal performance was for GMM6 system at stream weight of 0.8, giving 3.7% WER, a relative improvement of 10.9%.
- Streaming MFCC and PLP features gave little improvement.

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

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

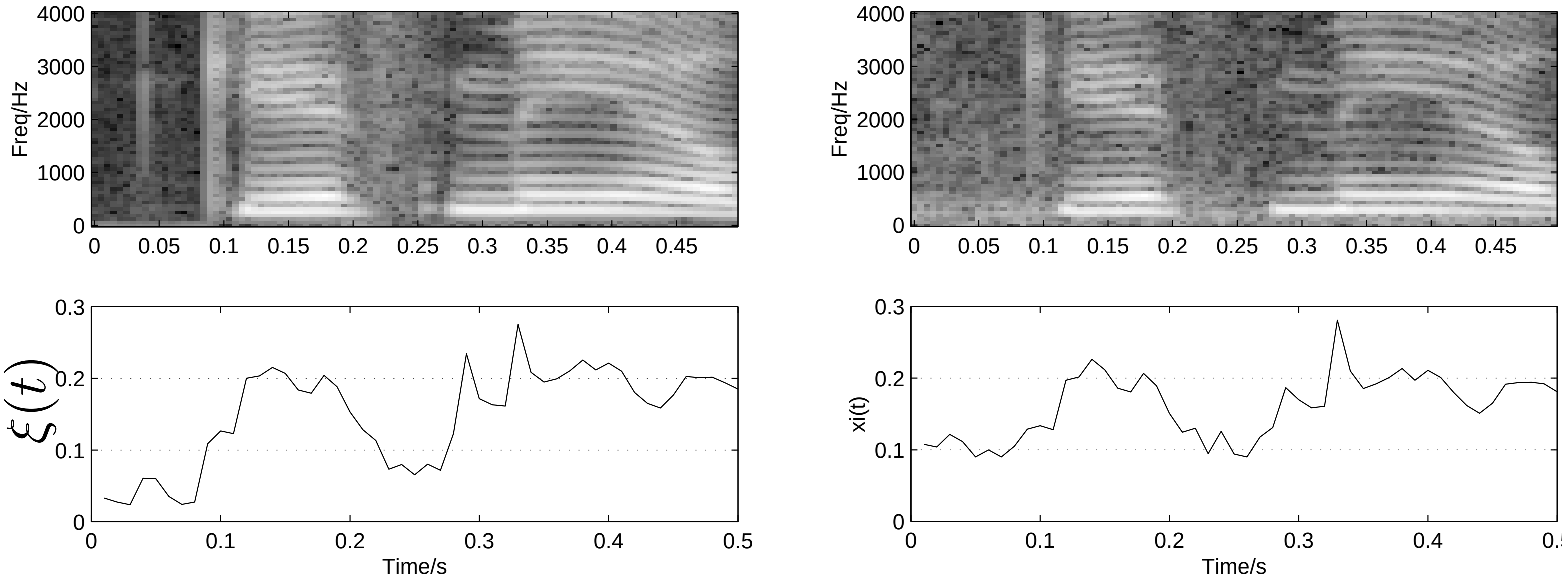
- Use standard synchronous stream system

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

- 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

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

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
GMM6	12.43
GMM6 feature mean normalisation	12.02

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

GMM Features in Noise

- Peak representations of speech are inheirently robust to some noise sources
- Noise sources with strong peak structures (ie background babble) can corrupt features significantly
- Unlike most peak representations, can reconstruct spectrum from GMM features
- Can compensate for noise at feature extraction stage by estimating clean speech parameters given noise model
- Alternatively can generate noise compensated model set given clean model set and noise model

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

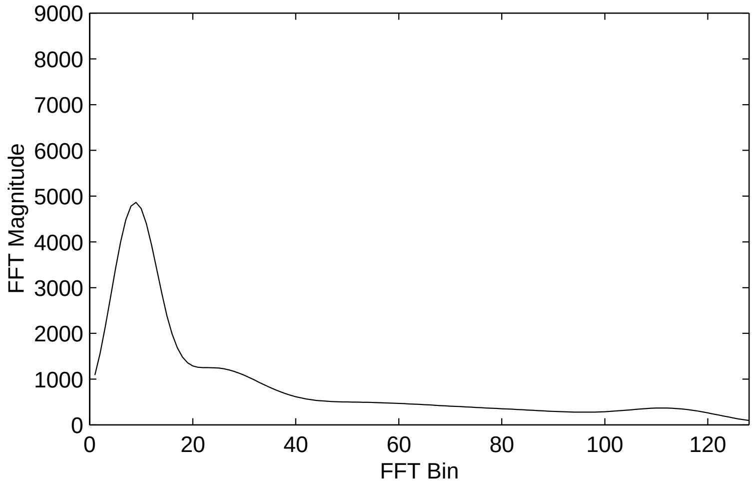
- Reconstructs spectra $\boldsymbol{x}_{jm}$ from GMM parameters of each state j and component m in model

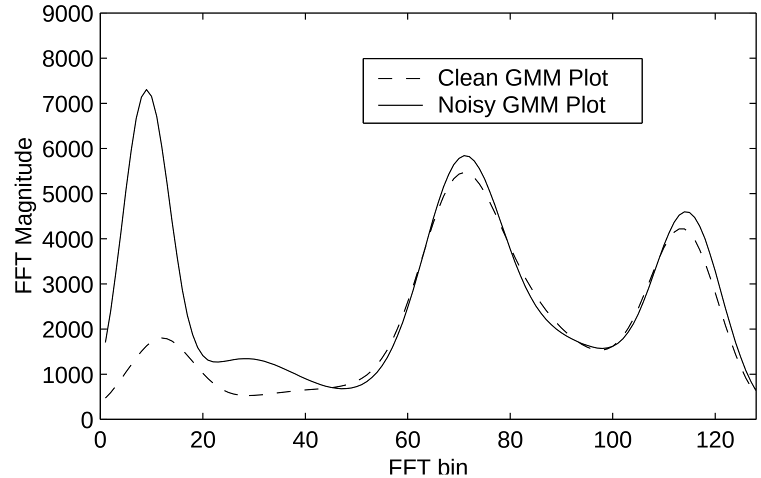
- Noise corrupted spectra is formed by adding spectra from noise spectrum $\boldsymbol{q}$

- Parameters for noisy data $\hat{\boldsymbol{\theta}}_{jm}$ are re-estimated

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

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

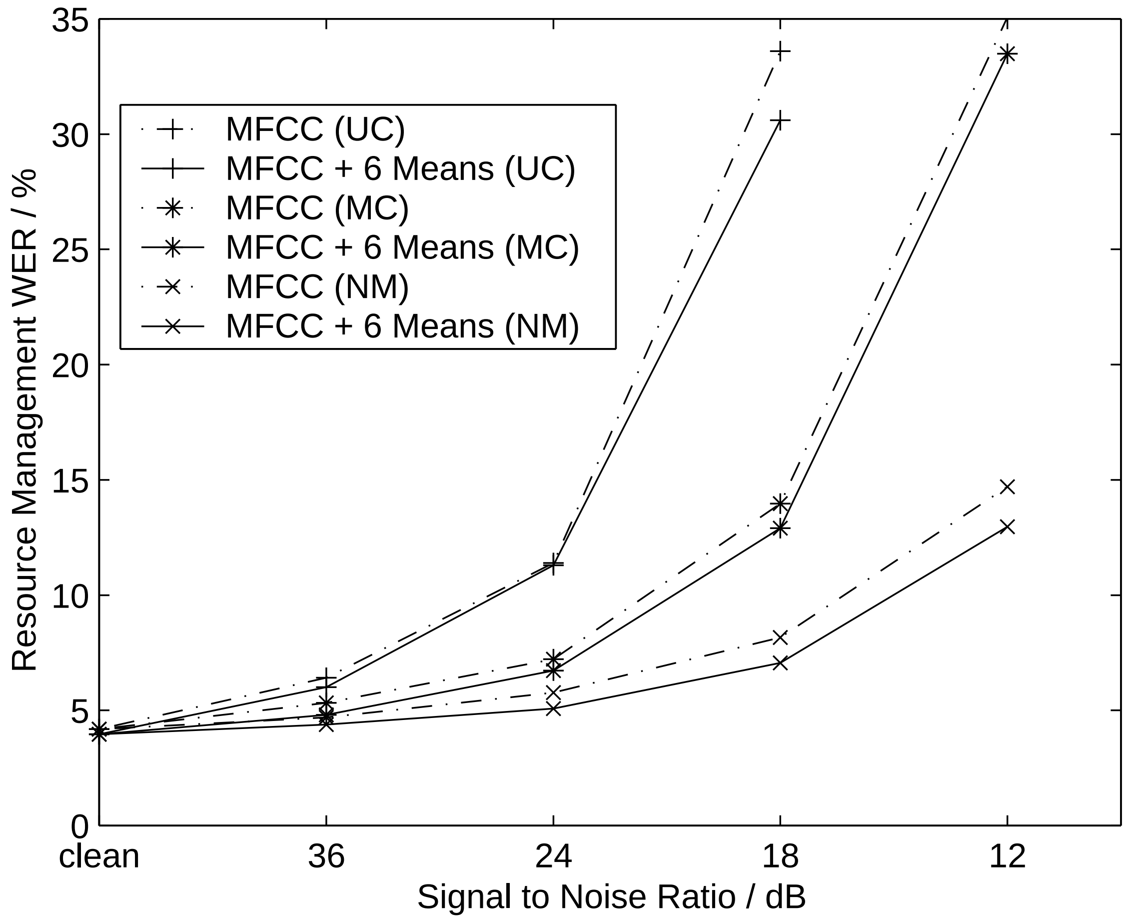
RM Results in additive noise - I

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

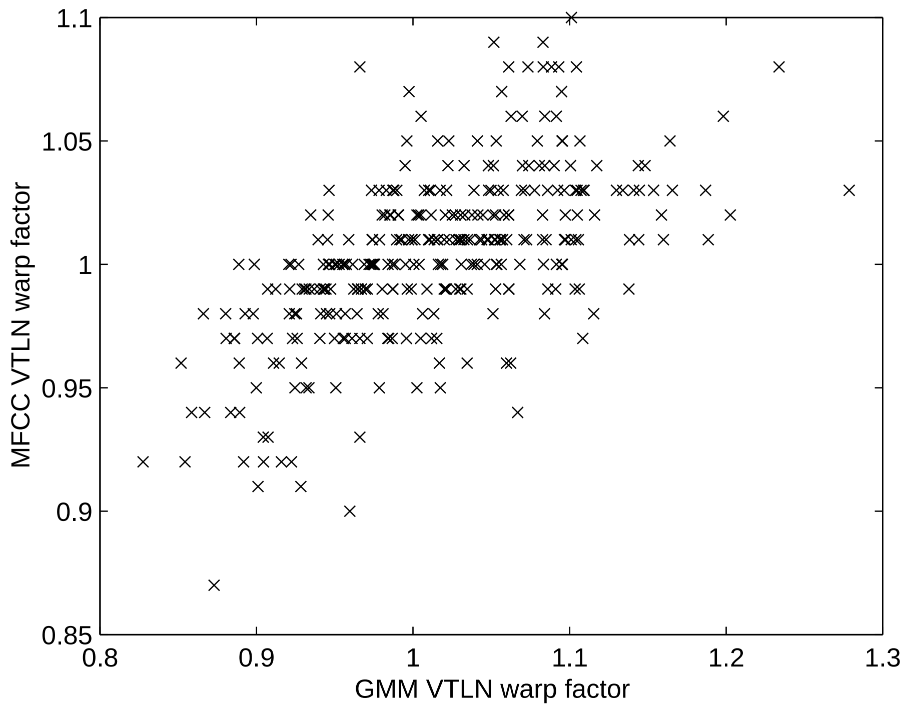


- Adding GMM features to MFCCs gives small improvements over a range of SNRs.

Speaker adaptation

- GMM features are directly represented in spectrum - position of compenent means are frequency bin values
- Can implement a VTLN approach by scaling the component means
- CMN approach approximates VTLN for GMM system
- Diagonal feature transforms will scale features for VTLN and spectral tilt effects.

Speaker adaptation



- Obtained an constrained diagonal MLLR transform for WSJ speakers
- Regression fit to GMM means warpings yields VTLN factors correlated to MFCC Brent estimated ML search parameters.

Unconstrained MLLR

- Adapting the data using a speech/silence full MLLR transform

Type of Transform	MFCC	MFCC + 6 Means	GMM6
None	9.75	9.56	12.0
UC MLLR	8.69	8.36	10.37
C MLLR	8.77	8.84	11.26
C MLLR + SAT	7.98	8.45	11.32

- 4% improvement incorporating GMM features with MFCCs and using UC MLLR
- Performance degrades when feature space transforms are used
- Systems using diagonal feature transforms did improve in CMLLR systems

Conclusions

- Fitting a GMM to speech provides features with information complementary to MFCC parameterisation.
- Incorporating GMM features with MFCCs by concantenating feature vectors reduces error rates on RM task.
- Combining MFCCs with GMM features using synchronous streams measure of confidence yields no significant improvement over concatenating into a single feature vector

Conclusions

- 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
- Relative improvements incorporating GMM features with an MFCC parameterisation are maintained with a MLLR adaptation
- GMM features are not suited to feature-space transforms and constrained MLLR approaches