

## Environmental robustness

Steve Renals

Automatic Speech Recognition— ASR Lecture 12  
4 March 2013

## Overview

### Today's lecture

- Recognising speech in presence of additive noise
- Feature compensation approaches
- Model compensation approaches

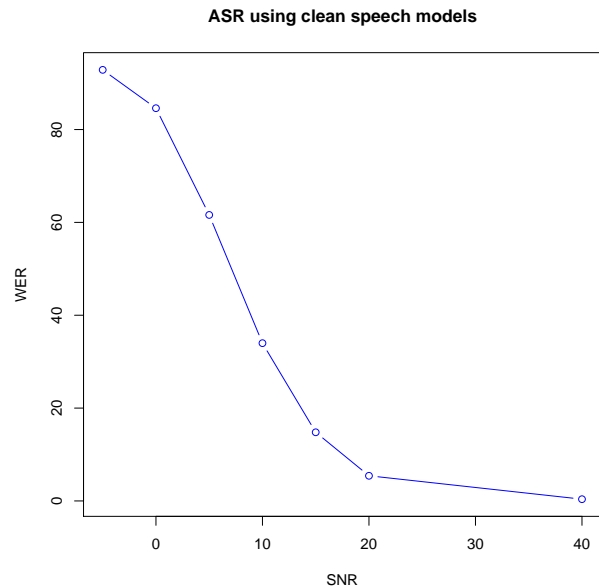
## Additive Noise

- Multiple acoustic sources are the norm rather than the exception
- From the point of view of trying to recognize a single stream of speech, this is additive noise
- **Stationary noise**: frequency spectrum does not change over time (e.g. air conditioning, car noise at constant speed)
- **Non-stationary noise**: time-dependent frequency spectrum (e.g. breaking glass, workshop noise, music, speech)
- Measure the noise level as SNR (signal-to-noise ratio), measured in dB
  - 30dB SNR sounds noise free
  - 0dB SNR has equal signal and noise energy

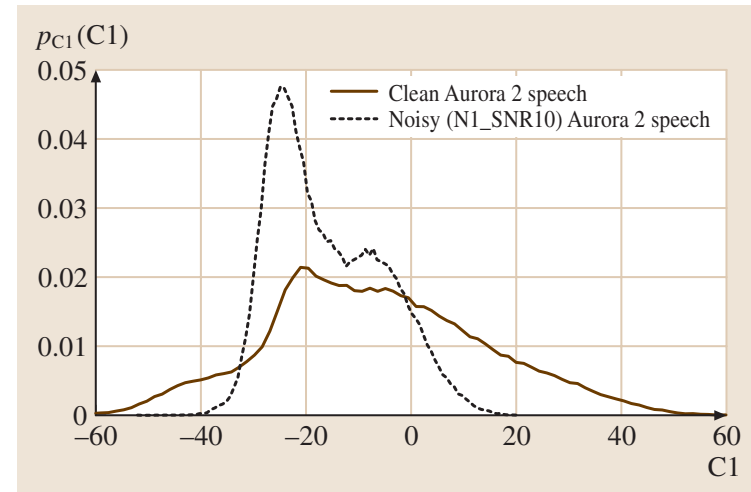
## Aurora-2

- Aurora is a standard set of speech + noise databases used in robust ASR research
- Aurora-2 speaker-independent continuously spoken strings of digits (TI-digits)
  - 11 word vocabulary
  - Artificially added noise of different types:
    - A: subway, babble, car exhibition
    - B: restaurant, street, airport, station
    - C: subway, street

## Recognizing Aurora-2 using Clean Speech Models



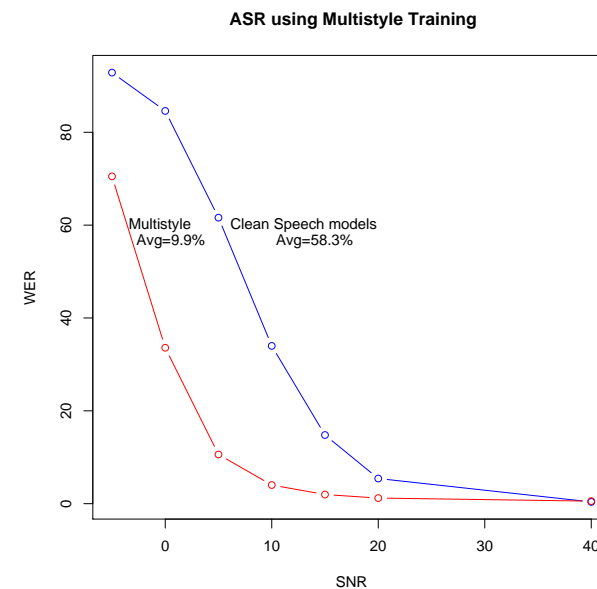
## Mismatch between clean and noisy speech



## Multistyle Training

- **Basic idea:** Don't train on clean speech, but train on speech with a similar noise level (and noise type)
- *Matched condition* — training in the same noise conditions as testing — is rarely possible since the test conditions are nearly always partly unknown
- *Multi-style training* — train with speech data in a variety of noise conditions
- It is possible to artificially mix recorded noise with clean speech at any desired SNR to create a multi-style training set
- Advantage: training data much better matched to test conditions
- Disadvantage: acoustic model components become less discriminative and less well matched to the training data
- Model adaptation — can further reduce errors using an adaptation technique such as MLLR

## Recognizing Aurora-2 using Multistyle Training



## Feature normalization

- **Basic idea:** Transform the features to reduce mismatch between training and test
- **Cepstral Mean Normalization (CMN):** subtract the mean of the feature vectors from each feature vector, so each feature vector element has a mean of 0
- CMN makes features robust to some linear filtering of the signal — adds robustness to varying microphones, telephone channels, etc.
- **Cepstral Variance Normalization (CVN):** Divide feature vector by standard deviation of feature vectors, so each feature vector element has a variance of 1
- Cepstral mean and variance normalisation, CMN/CVN:

$$\hat{\mathbf{x}}_i = \frac{\mathbf{x}_i - \boldsymbol{\mu}(\mathbf{x})}{\boldsymbol{\sigma}(\mathbf{x})}$$

On Aurora-2 CMN/CVN reduces multistyle training WER from 9.9% to 7.0%

## Feature compensation: Spectral subtraction

- **Basic idea:** Estimate the noise spectrum and subtract it from the observed spectra
- Any feature vector can then be computed from the noise-subtracted spectrum
- Problems:
  - Need to estimate noise spectrum from a period of non-speech: requires good speech/non-speech detection
  - Errors in the noise estimate (perhaps arising from speech/non-speech separation errors) result in over-/under-compensation of the spectrum
- Low computational cost, widely used in practice
- “ETSI advanced front end” uses spectral subtraction and CMN
  - 11.4% WER on Aurora-2 (clean models)
  - 6.8% WER on Aurora-2 (multistyle training)

## Feature compensation: SPLICE

- **Basic idea:** Predict the observed clean speech  $\mathbf{x}$  from the observed noisy speech  $\mathbf{y}$
- Estimate a joint mixture model for noisy and clean speech:

$$p(\mathbf{y}, \mathbf{x}) = \sum_k p(\mathbf{x}|\mathbf{y}, k)p(\mathbf{y}, k)$$

- $p(\mathbf{x}|\mathbf{y}, k)$  is a Gaussian component to predict the clean speech from the noisy speech:

$$p(\mathbf{x}|\mathbf{y}, k) = N(\mathbf{x}; \mathbf{A}\mathbf{y} + \mathbf{b}, \boldsymbol{\Sigma}_{xy})$$

- $p(\mathbf{y}, k)$  is a weighted Gaussian component

$$p(\mathbf{y}, k) = N(\mathbf{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)P(k)$$

- Train parameters from *stereo data*: simultaneous clean and noisy recordings
- Can use maximum likelihood or minimum mean square error objective function

## Model-based compensation

- **Basic idea:** use the detailed acoustic models in the recognizer as the basis of the compensation scheme
- Feature compensation approaches use an additional (simple) model of the speech signal—at best, a GMM
- Model-based compensation: combine the clean-speech models with a noise model to result in a model of noisy speech
- Results in taking the product of clean speech and noise components:  $M$  clean speech components and noise components result in  $MN$  noisy speech components
- High computational complexity for noise models more complex than a single Gaussian
- Two important approaches
  - Parallel model compensation (PMC)
  - Vector Taylor series (VTS) approximation

## Parallel model combination (PMC)

- **Basic idea:** Assume speech and noise is additive in spectral domain, so transform models from cepstral to spectral domain, compute noisy speech model statistics, transform back to cepstral domain
- Assume Gaussian noise model  $N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$
- PMC, using log-normal approximation:
  - 1 Compute speech and noise models in cepstral domain
  - 2 Map to spectral domain using inverse of the DCT and exponential
  - 3 Combine speech and noise parameters in spectral domain

$$\begin{aligned}\boldsymbol{\mu}_y^f &= \boldsymbol{\mu}_x^f + \boldsymbol{\mu}_n^f \\ \boldsymbol{\Sigma}_y^f &= \boldsymbol{\Sigma}_x^f + \boldsymbol{\Sigma}_n^f\end{aligned}$$

Even simpler approximation assumes  $\boldsymbol{\Sigma}_y = \boldsymbol{\Sigma}_x$

## Vector Taylor Series (VTS)

- **Basic idea:** Estimate noisy speech statistics as a Taylor series expansion about the means of the clean speech and noise
- Model the relationship between clean speech, noise and noisy speech as:

$$\mathbf{y} = \mathbf{x} + \mathbf{g}(\mathbf{n} - \mathbf{x})$$

$\mathbf{g}$  is a nonlinear function mapping signal to noise ratio to the difference between clean and noisy speech

- Approximate using a first-order Taylor series expansion around the clean speech and noise means  $(\boldsymbol{\mu}_x, \boldsymbol{\mu}_n)$

$$\mathbf{y} = \boldsymbol{\mu}_x + \mathbf{g}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_x) + \frac{\partial \mathbf{y}}{\partial \mathbf{x}}(\mathbf{x} - \boldsymbol{\mu}_x) + \frac{\partial \mathbf{y}}{\partial \mathbf{n}}(\mathbf{n} - \boldsymbol{\mu}_n)$$

- This results in expressions for  $\boldsymbol{\mu}_y$  and  $\boldsymbol{\Sigma}_y$  based on the statistics (mean and covariance) of clean speech and noise models
- VTS, CMN/CVN and multistyle training results in state of the art Aurora-2 results: 6.2% WER

## Missing feature approaches

- **Basic idea:** Assume each point in time-frequency plane is either reliable or unreliable evidence for the speech signal, and use this reliability to compute likelihoods
- Inspired by auditory scene analysis: each time-frequency point is dominated by energy from just one source
- Form a noise mask for those parts dominated by noise, and treat these as “missing” data for the speech
- Adjust the likelihood computation to take account of missing information
- Finding the noise mask:
  - Use SNR estimates
  - Use perceptual criteria (harmonics, common onset, etc.)
  - Train a classifier

## Summary

- Feature compensation: cepstral mean/variance normalisation, spectral subtraction, SPLICE
- Model compensation: parallel model compensation, missing feature approaches
- (Uncertainty decoding: use direct estimate of  $p(\mathbf{y} | \mathbf{x})$  in model compensation)