T-61.182 Robustness in Language and Speech Processing

Speech recognition in noisy environments: A survey

Yifan Gong

presented by Tapani Raiko

Feb 20, 2003

About the Paper

- Article published in Speech Communication in 1995 (31 pages)
- A survey of 250 publications divided to 3 categories:
  - Noise resistance
  - Speech enhancement
  - Model compensation for noise
- Focus: Mismatch in training and operating environments

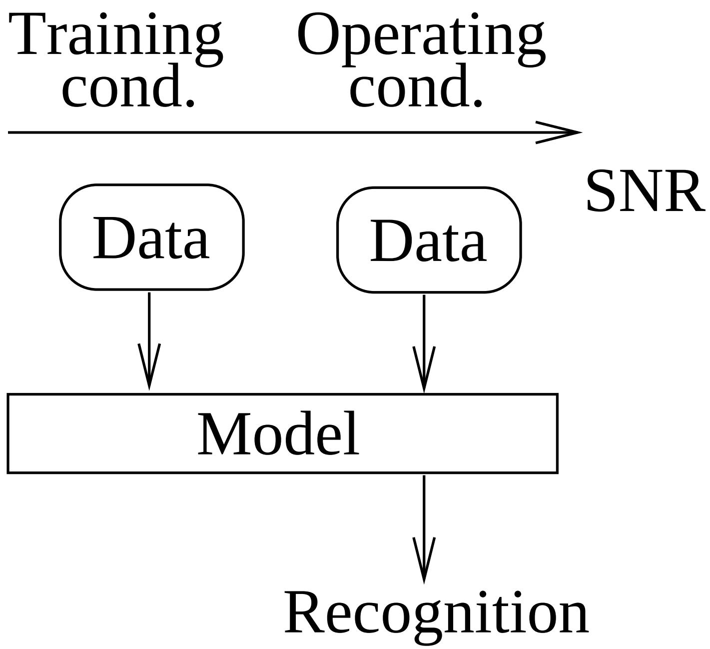
Introduction (1/2)

- Speech recognition in controlled situations has reached very high levels of performance
- Performance degrades in noisy situations
  - 100% to 30% accuracy in a car (90km/h)
  - 99% to 50% in a cafeteria
- Two phenomena:
  - Contaminated speech signal (typically additive, also convolutional)
  - Articulation variablity (called the Lombard effect)

Introduction (2/2)

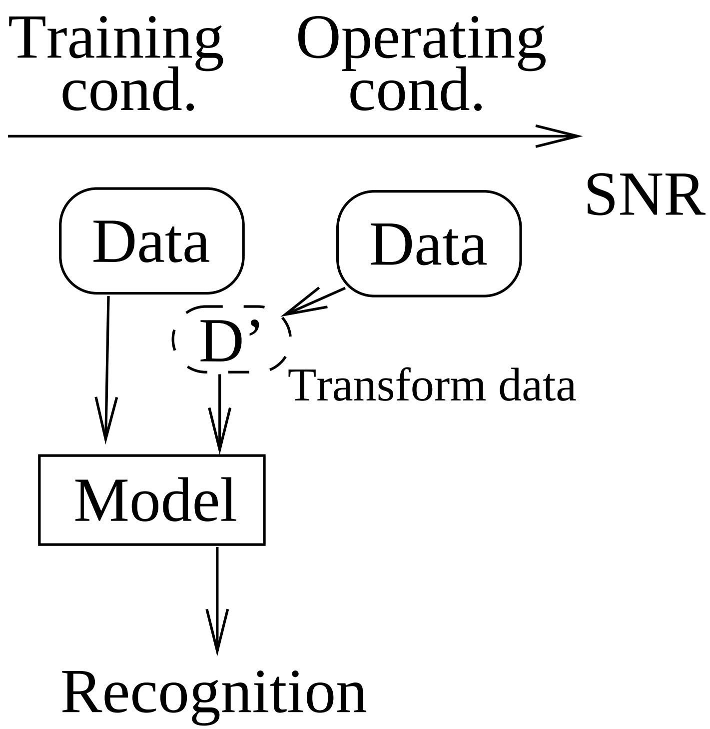
- A system trained with a given SNR performs worse in other SNR environments.
- What to do?
  - Search for noise resistant features and robust distance measures (1. Noise resistance)
  - Reduce the mismatch:
    * Remove noise from the signal (2. Speech enhancement)
    * Transform speech models to accommodate noise (3. Model compensation for noise)

1. Noise Resistance



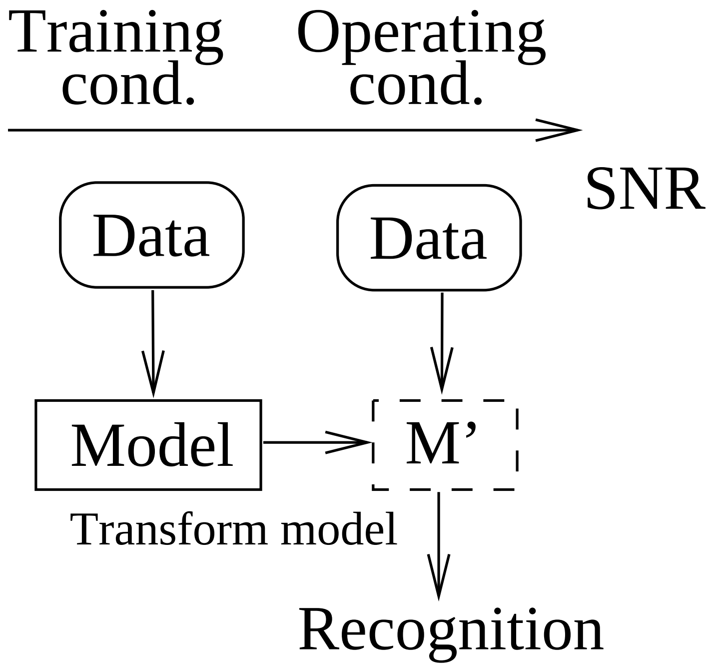
Search for noise resistant features and robust distance measures

2. Speech Enhancement



Remove noise from the signal

3. Model Compensation for Noise



Transform speech models to accommodate noise

1. Noise Resistance: Introduction

- Parameters of a recogniser are very sensitive to disturbances
- Focus on the effect of noise to the parameters
- Use derived feature parameters or similarity measurements (that are hopefully invariant to those effects)
- Weak or no assumptions about the noise
  - Both a strength and a weakness

1. Noise Resistance: Examples (1/2)

- Normalised cepstral vectors
  - Cepstrum is the Fourier transformation of spectrum
  - White noise corruption reduces the norm of cepstral vectors
  - The angle between ceptral vectors is less affected
- Spectral weightings methods (WLR, RPS, SWL)
  - Emphasize more spectral peaks than valleys
  - De-emphasize low quefrency terms of the cepstrum
- Multi-layer perceptron as phoneme classifier
  - Generalises better than e.g. k-nearest neighbour

1. Noise Resistance: Examples (2/2)

- Computational models of the auditory system for speech
  - Computationally expensive
  - Wavelet transform followed by a compressive nonlinearity
  - Frequency dependent lateral inhibition function
- Slow variation removal
  - Many noises vary slowly compared to speech
  - CMN: Remove the mean from cepstral vectors
  - Use time derivatives of cepstra
  - RASTA, RASTA-PLP, J-RASTA, J-RASTA-PLP
    (J handles also convolutional noise)

2. Speech Enhancement: Introduction

- A preprocessing step
- Developed for speech quality improvement
- Criteria are usually not related to recognition accuracy
- A priori information about the speech and the noise
- Enhanced SNR does not always improve recognition performance (the case with basic Wiener or Kalman filtering)

2. Speech Enhancement: Examples (1/2)

- Parameter mapping
  - Data: speech with and without noise
  - Teach a neural network to map noisy vectors to clean vectors
  - Highly dependent on training data
- Spectral subtraction
  - Estimate noise spectrum during non-speech periods
  - Subtract noise from the power spectrum
  - A special case of a Wiener filter
  - Noise masking is related (=ignore everything under a power threshold)

2. Speech Enhancement: Examples (2/2)

- Comb filtering
  - Estimate the period of the speech
  - Use only the corresponding frequency and its multiples
- Bayesian estimation
  - A generative model for latent true speech with noise
  - Estimate the posterior of the speech as the enhanced signal
  - Equivalent to template-based estimation (formulated without probabilities)

3. Model Compensation for Noise: Introduction

- Accept the precence of noise
- Hidden Markov Models (HMMs) as the framework
- Model parameteres are optimised during operation
- At very low SNRs problematic (at least in 1994)

3. Model Compensation for Noise: Examples (1/2)

- Decomposition of HMMs (PMC, STM)
  - $N \times M$ state HMM, N states for speech and M states for noise
  - Parts are trained separately
  - Assumes Gaussian distributions when actually at low SNR some are bimodal
- State-dependent Wiener filtering
  - Wiener filter uses the ratio of power spectrum of clean speech over the noisy speech
  - Power spectrum of speech is very non-stationary
  - Idea: HMMs automatically divide speech into quasi-stationary segments!

3. Model Compensation for Noise: Examples (2/2)

- Duration models
  - Duration structures of speech are less affected by noise
- Adaptation of HMMs
  - Train an HMM with lots of clean speech data
  - Use just a small amount of noisy speech to find a mapping of HMM parameters to the noisy environment
- Discriminative HMMs
  - Instead of maximum likelihood use maximum classification accuracy

Training Data Contamination

- Does not fit any of the three categories
- E.g. mix cafeteria recording to clean speech data
- Used as a benchmark
- Sensitive to noise level and type
- Cannot cope with the Lombard effect
- At equivalent SNRs, Gaussian noise is worst $\longrightarrow$ a lower bound

Conclusion (1/2)

- Focus: Mismatch between training and operating conditions
- Are properties of the noise known?
  Is computing power cheap?
  - No: Use 1. (feature-similarity-based)
  - Yes: Use 2. or 3. (transformation-based)
- Different techniques may be combined
- Non-stationary noise is a hot topic (1994)

Conclusion (2/2): Key Issues

- Accurate speech and noise models (state decomposition)
- Incorporating a dynamical model (HMM)
- Incorporating frequency correlations (LPC, SOM,. . .)
- Weighting portions of speech based on their SNR
- Class dependent processing (class $\in$ {word, phoneme, sound class, HMM state, VQ codebook vector})
- Optimisation criteria (discriminative training)
- Human auditory system as inspiration